Title

A doxastic behavioral characterization of generalized backward induction

Permalink
https://escholarship.org/uc/item/2sf6r7p5

Author
Bonanno, Giacomo

Publication Date
2014-11-01

DOI
10.1016/j.geb.2014.10.004

Peer reviewed

Published in Games and Economic Behavior, Volume 88, 2014, 221-241.
(http://dx.doi.org/10.1016/j.geb.2014.10.004)

A doxastic behavioral characterization of generalized backward induction [☆]

Giacomo Bonanno

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

Abstract

We investigate an extension of the notion of backward induction to dynamic games with imperfect information and provide a doxastic characterization of it. Extensions of the idea of backward induction were proposed by [1] and later by [2], who also provided a doxastic characterization in terms of the notion of common belief of future rationality. The 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 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 players' beliefs are modeled as temporal, rather than conditional, beliefs and rationality is defined in terms of actual choices, rather than hypothetical plans.

Keywords: extensive-form game, generalized backward induction, dynamic interactive beliefs, rationality, behavioral model
2000 MSC: 91A18, 91A40
JEL: C7

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] Recently [1] extended the notion of backward induction to extensive games with imperfect information: his "backwards (rationalizability) procedure" iteratively eliminates strategies and conditional belief vectors starting from the end of the game and proceeding backwards towards the root. Subsequently [2] provided an epistemic characterization of this procedure in terms of the notion of "common belief in future rationality".[2] [2] also introduced a new algorithm, the "backward dominance procedure", which

[☆] I am grateful to an Associate Editor and two anonymous reviewers for helpful comments and suggestions.

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

[1] For recent surveys of the literature see [3, 4].

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

differs from Penta's procedure in that it operates only on strategies, rather than strategies and conditional belief vectors,[3] and showed that the strategies that survive the backward dominance procedure are exactly the strategies that can rationally be chosen under common belief in future rationality if one does not impose (common belief in) Bayesian updating.

The purpose of this paper is to offer not a novel solution concept but a novel analysis of the notion of backward induction in games with possibly imperfect information using dynamic behavioral models where players' beliefs are temporal, rather than conditional, beliefs and strategies plays no role. We introduce a backward-induction procedure that operates on *choices* rather than strategies (we call it "generalized backward induction" in order to distinguish it from the procedures introduced by [1] and [2]) and whose output is a set of *terminal histories*, rather than a set of strategy profiles. Furthermore, we provide a doxastic characterization of it in terms of the notion of "forward belief in rationality", which is conceptually very similar to Perea's notion of common belief in future rationality. The novelty of this contribution lies in showing that strategies and (objective or subjective) counterfactuals are not needed in order to obtain a doxastic foundation for a general notion of backward induction.

The epistemic models of dynamic games used in the literature[4] are static structures that postulate, for each player, a complex set of conditional belief hierarchies. In these models, at every information set of hers 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 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." ([10], 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 [11] introduced a simpler kind of models that are explicitly dynamic and make no use of counterfactuals or dispositional belief revision; furthermore, these are "behavioral" models in which states are described in terms of the actual choices made by the players rather than in terms of hypothetical plans.[5] In these models there are no complex epistemic states (consisting of full belief revision policies): only the actual beliefs of a player when it is her turn to move.

We extend the epistemic models of [11] to games with imperfect information (allowing for the possibility of simultaneous moves). 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. If h is the decision history reached in state ω and time t and i is an active player there, then player i has to choose an action from the set $A_i(h)$ 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

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

[4] See, for example, [5, 6, 7, 2, 8, 9].

[5] Behavioral models were introduced by [12].

(2) what will happen if she chooses action a, for every $a \in A_i(h)$. These beliefs can then be 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" ([13]): 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 doxastic 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 players (1) are rational, (2) believe that at future dates the active players will be rational, (3) believe that at future dates the active players will believe that future players will be rational, (4) believe that at future dates the active players will believe that future players will believe that future players will be rational, and so on. Call this event **FBR**.[6] We show (Proposition 1) that, in an arbitrary model of a game, if ω is a state such that $\omega \in \mathbf{FBR}$ 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 (possibly iteratively) choices that are strictly dominated there and then iterates backwards towards the root. We restrict attention to a class of games that is essentially the same as the class of games considered by [2].[7] 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 $\omega \in \mathbf{FBR}$ and the terminal history associated with ω is z. Thus 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, while Sections 3 and 4 contain the definition of rationality and provide a doxastic characterization of the output of the generalized backward induction algorithm. In order to lighten the exposition, in Section 3 attention is restricted to games without simultaneous moves and Section 4 details the additions that need to be made for games that have simultaneous moves. Since the word 'epistemic' refers to knowledge, while we deal with the more general notion of - possibly erroneous - belief, we use the expression 'doxastic characterization' rather than 'epistemic characterization'. Indeed, unlike the condition provided in [11] 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 5 concludes with a discussion of the proposed approach and of relevant literature. The proofs are given in the Appendix.

2. State-time representation of extensive games

We use the history-based definition of extensive-form game (see, for example, [14]). 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

[6] When simultaneous moves are present the description of the event **FBR** includes also common belief at time t that all active players at time t are rational and that future players will be rational.

[7] [2] considers games where there is an unambiguous ordering of the information sets, while we restrict attention to von Neumann games, where decision histories that belong to the same information set have the same length. As argued in Footnote 10, the two classes of games are essentially equivalent.

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* and possibly simultaneous moves is given by the following elements:

- A finite set N of players.

- For every player $i \in N$ a finite set of actions A_i. If $J \subseteq N$ is a non-empty subset of players, let $A_J = \prod_{i \in J} A_i$ be the action profiles for players in J and let $A = \bigcup\limits_{\varnothing \neq J \subseteq N} A_J$ be the set of action profiles over all possible non-empty subsets of N. If $a \in A$ contains only one action (that is, $a \in A_i$ for some $i \in N$), we call a a *simple move*, otherwise we call it a *simultaneous move*.

- 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.

- A function $\iota : D \to 2^N \backslash \varnothing$ that assigns to each decision history h a non-empty set of players who move, or *are active*, at history h. For every $i \in N$, let $D_i = \{h \in D : i \in \iota(h)\}$ be the set of histories at which player i is active. For every history $h \in D_i$, we denote by $A_i(h)$ the set of actions available to player i at h, that is, $a_i \in A_i(h)$ if $a_i \in A_i$ and there is a profile of actions $a \in A$ to which a_i belongs such that $ha \in H$; furthermore, we denote by $A(h)$ the set of action profiles for the active players at h, that is, $A(h) = \prod_{i \in \iota(h)} A_i(h)$.

- For every player $i \in N$, an equivalence relation $\approx_i$ 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 $A_i(h') = A_i(h)$, 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.[8]

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

The top part of Figure 1 shows two alternative representations of the following extensive-form: $H = \{\emptyset, b, a, a(c, e), a(c, f), a(c, g), a(d, e), a(d, f), a(d, g)\}$, $\iota(\emptyset) = \{1\}, \iota(a) = \{1, 2\}, D_1 = \{\emptyset, a\}, D_2 = \{a\}, A_1(a) = \{c, d\}, A_2(a) = \{e, f, g\}$, etc.

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 $z \succsim_i z'$.

[8] 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.

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.[9]

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 action profiles that appear in h; for example, if $h = a_1(a_2, a_3, a_4)a_5(a_6, a_7)$ then $\ell(h) = 4$. 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 $0 \leq t \leq \ell(h)$, we denote by h_t the prefix of h of length t. For example, if $h = a_1(a_2, a_3, a_4)a_5(a_6, a_7)$, then $h_0 = \emptyset$, $h_2 = a_1(a_2, a_3, a_4)$, etc.

We follow [2] and restrict attention to extensive forms where there is an unambiguous ordering of the information sets. However, for notational simplicity, it will be convenient to restrict attention to the essentially equivalent class of von Neumann extensive forms, which are defined as follows.[10]

Definition 1. An extensive form is *von Neumann* if any two decision histories that belong to the same information set have the same length: $\forall i \in N$, $\forall h, h' \in D_i$, if $h' \in I_i(h)$ then $\ell(h) = \ell(h')$.

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*.

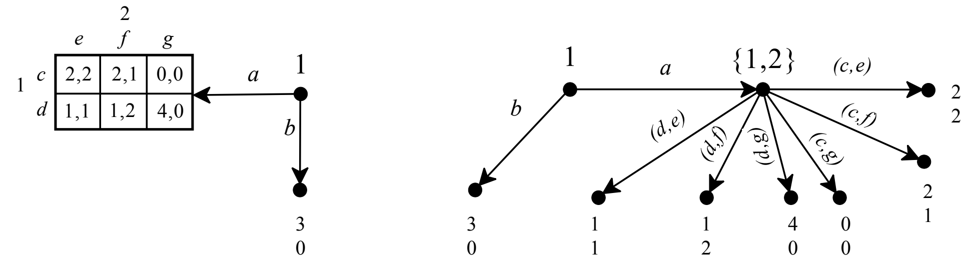
Definition 2. Given a von Neumann extensive form, a *state-time representation* of it is a triple $\langle \Omega, T, \zeta \rangle$ where Ω is a set of states, $T = \{0, 1, ..., m\}$ is a set of instants with $m \geq \ell^{\max} - 1$ (recall that $\ell^{\max}$ is the depth of the game) and $\zeta : \Omega \to Z$ is a function that assigns to every state a terminal history. Given a state-instant pair $(\omega, t) \in \Omega \times T$, define

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

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 (possibly simultaneous) 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"

[9] 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 her own preferences, it is much more demanding to assume that she knows the preferences of her 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).

[10] Extensive forms where there is an unambiguous ordering of the information sets can be transformed into von Neumann extensive forms in a trivial way by adding, along some histories, a fictitious player (the "clock") who always has singleton information sets and only one choice (moving the clock forward) at each history assigned to it.



state ω :	α	β	γ	δ	ε	η	θ
ζ(ω) :	*b*	*a(d,e)*	*a(d,f)*	*a(d,g)*	*a(c,g)*	*a(c,f)*	*a(c,e)*
time:							
0	∅	∅	∅	∅	∅	∅	∅
1	*b*	*a*	*a*	*a*	*a*	*a*	*a*
2	*b*	*a(d,e)*	*a(d,f)*	*a(d,g)*	*a(c,g)*	*a(c,f)*	*a(c,e)*

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

at state ω, that is, the partial history of 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 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, $\zeta_1(\alpha) = \zeta_2(\alpha) = b$, $\zeta_1(\beta) = a$, $\zeta_2(\beta) = a(d, e)$, etc.

We want to define the notion of rational *behavior* in a game (as captured by choices actually made), rather than rational planning (as captured by strategies), and examine its implications. In order to do so we will introduce the notion of a dynamic model of a game. First 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, $\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega\mathcal{B}_i\omega'\}$.[11] $\mathcal{B}_i$ 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: $B_iE = \{\omega \in \Omega : \mathcal{B}_i(\omega) \subseteq E\}$. Hence B_iE is the event that individual i believes E.[12] 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 : $B_iE \subseteq \neg B_i \neg E$, where,

[11] 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.

[12] In a probabilistic setting the interpretation of the event B_iE would be "the set of states where player i attaches probability 1 to event E".

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 : $B_iE \subseteq B_iB_iE$) and euclideanness corresponds to negative introspection (if the individual does not believe E then she believes that she does not believe E: $\neg B_iE \subseteq B_i\neg B_iE$) (for more details see [15]).

In order to lighten the exposition, in the next section we will concentrate on extensive games without simultaneous moves and then, in Section 4, explain the additions that need to be made in order to accommodate simultaneous moves.

3. Games without simultaneous moves

In this section we restrict attention to extensive games where, at each decision history, exactly one player is active.

We say that 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 $A_i(h)$. 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 $a \in A_i(h)$. 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.[13] 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.

3.1. Dynamic models

Definition 3. Given a von Neumann extensive form, a *dynamic model* of it 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 the extensive form (see Definition 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:

[13] 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.

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.[14]
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 $A_i(h)$.
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.[15]
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 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 $h'a$. 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. 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 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.[16] 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; in other words, in her mind each of her currently available actions is still a

[14] 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.

[15] 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.

[16] The role of counterfactuals in the standard approach is discussed in details in [16].

"live possibility".[17] 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.

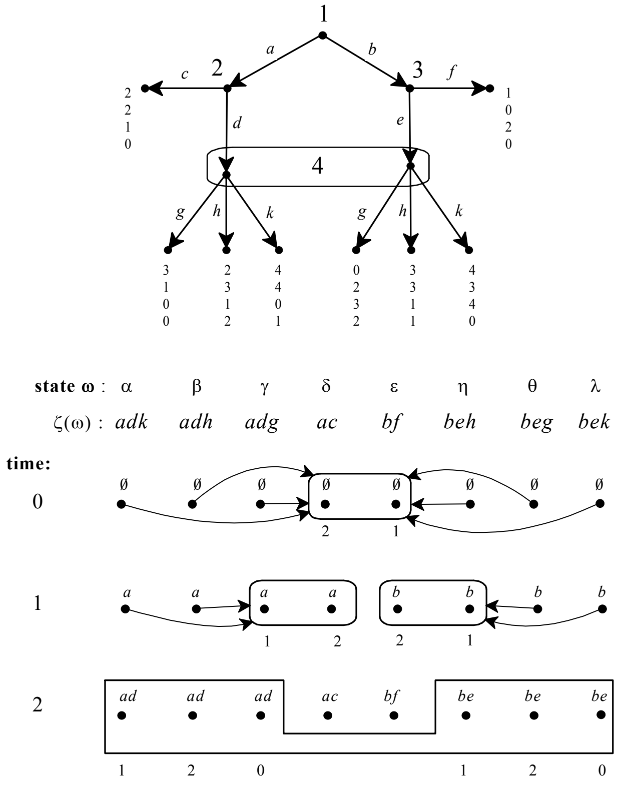


Figure 2: An von Neumann extensive game and a model of it.

Figure 2 shows a game and 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 rectangle or there is an arrow from ω to the rectangle containing ω'.[18] The relations shown in Figure 2 are

[17] 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 her forthcoming choice, the player is allowed to have (possibly erroneous) beliefs about what choice she will make at a later time. For an extensive discussion of this point see Section 4 in [11].

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

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, 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 ($\mathcal{B}_{4,2}$).[19] 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 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 states that she considers possible.

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; in other words, we deal with belief rather than knowledge).

Remark 3. Note that Definition 3 allows for "causally flawed" beliefs. For example, consider a game where Player 1 moves first, choosing between a and b, and Player 2 moves second and chooses between c and *d without being informed of Player 1's choice*. Definition 3 allows Player 1 to have the following beliefs at time 0: "if I play a, Player 2 will play c, while if I play b then he will play d". Such beliefs can be considered "implausible" or "irrational" on the grounds that there cannot be a causal link between Player 1's move and Player 2's choice, since Player 2 does not get to observe Player 1's move; thus a "causally correct" belief for Player 1 would require that the predicted choice of Player 2 be the same, no matter what Player 1 does. When the game does not have simultaneous moves, this restriction on beliefs is not needed for our results. We will, however, need to rule out causally flawed beliefs within the context of a simultaneous move (see Definition 6).

3.2. Rationality of choice

We shall use a very weak notion of rationality, which has been referred to in the literature as "material rationality" (see, for example, [17, 13, 18, 12]). 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.

[19] 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, $\mathcal{B}_{i,t}(\omega) = \varnothing$.

Definition 4. Fix a state-instant pair (ω, t) and suppose that $\zeta_t(\omega)$ is a decision history of player i. Let $a, b \in A_i(\zeta_t(\omega))$ 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 3, she must consider it possible that she takes any of her available actions at h'].

Definition 5. 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 $b \in A_i(\zeta_t(\omega))$, it is not the case that player i believes at (ω, t) that b is better than a (see Definition 4).

For example, in the model of Figure 2, Player 1 is rational at date 0 and states α, β, γ and δ (but not at the remaining states), 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 and believes that h is better than k; in fact, k is strictly dominated by 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 $\mathbf{R}_{i,t} \subseteq \Omega$ the event that (that is, the set of states at which) player i is rational at date t and let $\mathbf{R}_t = \bigcap_{i \in N} \mathbf{R}_{i,t}$. Since, by definition, every player is rational at a state-instant pair where she is not active, $\mathbf{R}_t$ can be described as the event that *the active player* (if there is one) *is rational at date t*. 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) (see Definition 5). 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) (for example, in the model of Figure 2, the active player at $(\delta, 1)$ is Player 2, while the active player at $(\epsilon, 1)$ is Player 3). In the model of Figure 2 we have that $\mathbf{R}_0 = \{\alpha, \beta, \gamma, \delta\}$, $\mathbf{R}_1 = \{\delta, \epsilon\}$ 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 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 $B_t E = \bigcap_{i \in N} B_{i,t}E$ (thus $\omega \in B_t E$ if and only if $\bigcup_{i \in N} \mathcal{B}_{i,t}(\omega) \subseteq E$). Then $B_t E$ is the event that "the active player believes E at time t" (which, by Condition 1 of Definition 3, 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 $B_t E = \bigcap_{i \in N} B_{i,t}E$ (thus $\omega \in B_t E$ if and only if $\mathcal{B}_t(\omega) \subseteq E$). It follows that if i is the active player at $\zeta_t(\omega)$, then $\mathcal{B}_t(\omega) = \mathcal{B}_{i,t}(\omega)$ and, for every event E, $\omega \in B_t E$ if and only if $\omega \in B_{i,t}E$ if and only if $\mathcal{B}_{i,t}(\omega) \subseteq E$.

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 5). 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 not only about the future choices but also about the future beliefs of other players (even about her own future choice(s) and beliefs, as is the case at state ϵ and date 0 in the model of Figure 5 discussed in Section 5.2).

3.3. 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:[20]

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

... 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 5, 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 ($r \geq 1$, $t_r < k$).

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

$$\mathbf{FBR}_{m-1} = \mathbf{R}_{m-1},$$

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

[20] 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 $\ell^{max} = 4$ $\mathbf{FBR} = \mathbf{R}_0 \cap \mathbf{R}_1 \cap \mathbf{R}_2 \cap \mathbf{R}_3 \cap B_0(\mathbf{R}_1 \cap \mathbf{R}_2 \cap \mathbf{R}_3) \cap B_1(\mathbf{R}_2 \cap \mathbf{R}_3) \cap B_2\mathbf{R}_3 \cap B_0B_1(\mathbf{R}_2 \cap \mathbf{R}_3) \cap B_0B_2\mathbf{R}_3 \cap B_1B_2\mathbf{R}_3 \cap B_0B_1B_2\mathbf{R}_3$.

Thus, for example, $\mathbf{FBR}_{m-2} = \mathbf{R}_{m-2} \cap B_{m-2}\mathbf{R}_{m-1} \cap \mathbf{R}_{m-1}$ and
$\mathbf{FBR}_{m-3} = \mathbf{R}_{m-3} \cap B_{m-3}(\mathbf{R}_{m-2} \cap B_{m-2}\mathbf{R}_{m-1} \cap \mathbf{R}_{m-1}) \cap (\mathbf{R}_{m-2} \cap B_{m-2}\mathbf{R}_{m-1} \cap \mathbf{R}_{m-1})$.[21]

Finally, define

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

For example, 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 $\mathbf{R}_0 = \Omega, \mathbf{R}_1 = \{\gamma, \delta, \epsilon\}, \mathbf{R}_2 = \{\beta, \gamma, \epsilon\}, B_0\mathbf{R}_1 = B_1\mathbf{R}_2 = B_0B_1\mathbf{R}_2 = \Omega$ 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 from the set of states that she considers possible), 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 Corollary 1 below, this is not a coincidence.

Next we introduce an algorithm that, for every von Neumann extensive-form game without simultaneous moves, 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 (if the output of backward induction is thought of as a terminal history rather than a strategy profile). The procedure 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 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 $D(h)$. The set $D(h)$ is defined recursively as follows, letting $i = \iota(h)$ (recall that $\iota(h)$ is the player who moves at h):

1. If $\ell(h) = \ell^{max} - 1$ then $a \in D(h)$ if and only if $a \in A_i(h)$ and there exists a $b \in A_i(h)$ 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 that 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 that information set].
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 $a \in A_i(h)$ and there exists a $b \in A_i(h)$ such that, for every $h' \in I_i(h)$, the following holds: $u_i(z') < u_i(z'')$ for every $z', z'' \in Z$ such that $z' = h'aa_1 \dots a_p$ ($p \geq 0$), $z'' = h'bb_1 \dots b_q$ ($q \geq 0$), $a_1 \notin D(h'a)$, $b_1 \notin D(h'b)$ and, $\forall j = 2, \dots, p, \forall k = 2, \dots, q,\ a_j \notin D(h'aa_1 \dots a_{j-1})$ and $b_k \notin D(h'bb_1 \dots b_{k-1})$ [that is, if there exists another choice b that guarantees a higher utility than a at the information set that contains h, assuming that only undominated actions are played after the choice there].

Next we define the following function $f_{BI} : H \to 2^Z$ (2^Z is the set of subsets of the set Z of terminal histories): (1) if $h \in Z$ then $f_{BI}(h) = \{h\}$ and (2) if h is a decision history then

$$f_{BI}(h) = \{z \in Z : z = ha_1a_2...a_m,\ a_1 \notin D(h)\ and\ a_j \notin D(ha_1...a_{j-1}), \forall j = 2, ..., m\}. \tag{2}$$

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

[21] 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 $\mathbf{FBR}_0 = \varnothing$.

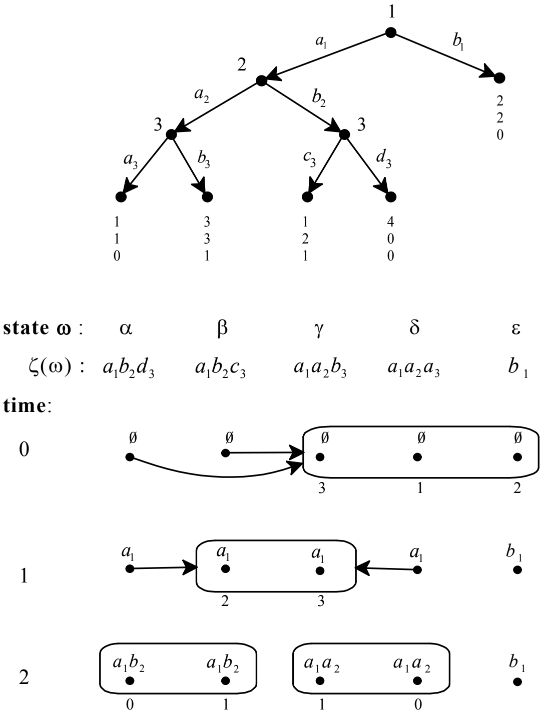


Figure 3: A perfect-information game and a model of it.

Finally define the event **BI** ⊆ Z as follows:

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

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 1. In the game of Part A of Figure 4, $D(b) = D(\emptyset) = \varnothing$ and thus $f_{BI}(b) = \{bc, bd\}$ and $\mathbf{BI} = f_{BI}(\emptyset) = \{a, bc, bd\}$. In the game of Part B of Figure 4, $D(a) = \{c\}$, $D(b) = \{f\}$, $D(\emptyset) = \{b\}$ and thus $f_{BI}(a) = \{ad\}$, $f_{BI}(b) = \{be\}$ and $\mathbf{BI} = f_{BI}(\emptyset) = \{ad\}$. 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 $\mathbf{BI} = f_{BI}(\emptyset) = \{a, bds, bdt\}$.[22]

[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 $f_{BI}(a_1b_2) = \{a_1b_2c_3\}$.

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) $\zeta(\omega) \in f_{BI}(\zeta_{t+1}(\omega))$.

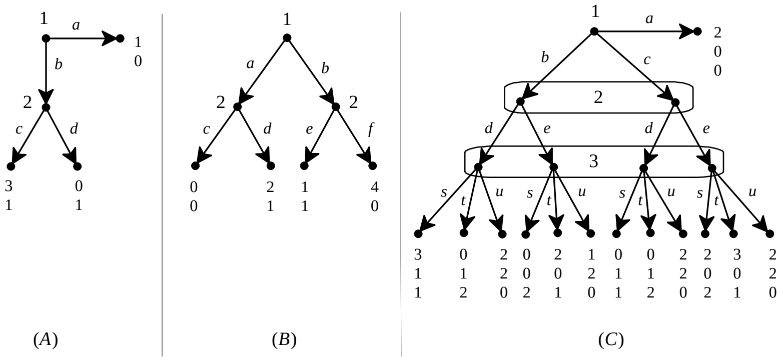


Figure 4: Three extensive games.

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 for the more general case where simultaneous moves are allowed (see Section 4).

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 $\zeta(\omega) = z$.*

An extensive game has *perfect information* if and only if every information set is a singleton, that is, if $h \in D_i$ then $I_i(h) = \{h\}$. A perfect-information game has *no relevant ties* if, $\forall i \in N$, $\forall h \in D_i$, $\forall a, a' \in A_i(h)$ 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]

4. Games with simultaneous moves

We now explain how the analysis can be extended to include games with simultaneous moves.

[23] [11] 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 [11] is stronger than **FBR** and implies it).

4.1. *Definition of model*

As anticipated in Remark 3, when several players move at a decision history h we need to add a further condition to Definition 3 which rules out "causally flawed" beliefs. First some notation. Let h be a decision history at which two or more players are active. Let i be one of the active players at h (that is, $i \in \iota(h)$) and let $a \in A(h)$ be an action profile at h. Then we write a as (a_i, a_{-i}) where a_i is player i's action in the profile a and a_{-i} is the profile of actions of the remaining players in $\iota(h)$.

Definition 6. Given a von Neumann extensive form (possibly with simultaneous moves), a *dynamic model* of it is a tuple $\left\langle \Omega, T, \zeta, \{\mathcal{B}_{i,t}\}_{i\in N, t\in T} \right\rangle$ that satisfies the properties of Definition 3 with Condition 2.3 replaced by the following:

2.3s If $i \in \iota(\zeta_t(\omega))$ (player i is one of the active players at $\zeta_t(\omega)$),
$\omega' \in \mathcal{B}_{i,t}(\omega)$, $(a_i, a_{-i}) \in A(\zeta_t(\omega'))$ and $\zeta_{t+1}(\omega') = \zeta_t(\omega')(a_i, a_{-i})$,
then, for every $b_i \in A_i(\zeta_t(\omega'))$, there is an $\tilde{\omega} \in \mathcal{B}_{i,t}(\omega)$ such
that $\zeta_{t+1}(\tilde{\omega}) = \zeta_t(\omega')(b_i, a_{-i})$.

Condition 2.3s (besides imposing the restriction of Condition 2.3 of Definition 3, namely that the player consider every action of hers a "live possiblity") rules out beliefs where player i considers it possible that if she plays a_i then the other players will simultaneously play a_{-i}, while if she plays some other action b_i then the other players will *not* play a_{-i}. As explained in Remark 3 such a belief would imply an implausible causal link between i's choice and the choices of her opponents. Without the restriction on beliefs given by Condition 2.3s, one could not rule out as irrational the play of a strictly dominated choice in a simultaneous game.[24]

Note that, at a history $h = \zeta_t(\omega)$ where player i is the only active player, Condition 2.3s coincides with 2.3 of Definition 3, since $A_{-i}(h) = \varnothing$. Hence Definition 6 reduces to Definition 3 in games where there are no simultaneous moves.

4.2. *Definition of rational choice*

Concerning the rationality of choice, Definition 4 needs to be reworded slightly in order to reflect the possible simultaneity of moves,[25] while Definition 5 remains unchanged.

4.3. *Definition of forward belief of rationality*

Concerning the notion of forward belief of rationality, we need to replace the individual belief operators with the common belief operator. Intuitively, an event E is commonly believed if everybody believes E and everybody believes that everybody believes E, and so on, *ad infinitum*. Accordingly, we define the common belief operator at time t, denoted by $\mathbb{CB}_t$, as follows. Let N^* be the set of non-empty finite sequences $< i_1, ..., i_m >$ ($m \geq 1$) in N (the set of players). Then, for every event $E \subseteq \Omega$,

$$\mathbb{CB}_t E = \bigcap_{<i_1,...,i_m>\in N^*} B_{i_1,t} B_{i_2,t} ... B_{i_m,t} E. \tag{3}$$

Thus, for example, $\mathbb{CB}_0 \mathbf{R}_1$ is the event that (that is, the set of states at which) it is common belief at date 0 that the active players at date 1 are rational.

[24] For example, it would be rational to "cooperate" in the simultaneous one-shot Prisoner's Dilemma game if one held the following beliefs: "if I cooperate she will cooperate and if I defect she will defect".

[25] As follows: fix a state-instant pair (ω, t) and let $i \in \iota(\zeta_t(\omega))$ and $a_i, b_i \in A_i(\zeta_t(\omega))$. We say that at (ω, t) *player i believes that b_i is better than a_i* if $\forall \omega_1, \omega_2 \in \mathcal{B}_{i,t}(\omega)$ such that $\zeta_t(\omega_1) = \zeta_t(\omega_2)$ and $\forall a_{-i} \in A_{-i}(\zeta_t(\omega_1))$ if $\zeta_{t+1}(\omega_1) = \zeta_t(\omega_1)(a_i, a_{-i})$ and $\zeta_{t+1}(\omega_2) = \zeta_t(\omega_2)(b_i, a_{-i})$, then $u_i(\zeta(\omega_1)) < u_i(\zeta(\omega_2))$.

Remark 7. Let $\mathcal{B}_{*,t}$ be the transitive closure of $\bigcup\limits_{i \in N} \mathcal{B}_{i,t}$. It is well-known that the following are equivalent: $\forall \omega \in \Omega, \forall E \in 2^{\Omega}$,

1. $\omega \in \mathbb{CB}_t E$,
2. $\mathcal{B}_{*,t}(\omega) \subseteq E$,
3. $\omega_m \in E$ for every sequence $\langle \omega_0, ..., \omega_m \rangle$ in Ω and $\langle i_1, ..., i_m \rangle$ in N ($m \geq 1$) such that $\omega_0 = \omega$ and, $\forall k = 1, ..., m,\ \omega_k \in \mathcal{B}_{i_k,t}(\omega_{k-1})$.

Furthermore, for every $i \in N$, $\mathbb{CB}_t E \subseteq B_{i,t}\mathbb{CB}_t E$.

The following remark is a consequence of the fact that inactive players believe everything (see Definition 3).

Remark 8. In games where there are no simultaneous moves $\mathbb{CB}_t E = B_t E$ for every event E and date t.

While in games without simultaneous moves it is sufficient to postulate that the active players are rational at every date t, when there are simultaneous moves it is necessary to add the assumption that "there is common belief among the active players at time t that they are all rational". This assumption cannot be expressed by the event $\mathbb{CB}_t\mathbf{R}_t$, because of the way in which we have modeled a player's belief at the time of choice: it cannot be postulated that an active player at date t believes that she is rational at date t (see Remark 2 and Footnote 17), while such a belief would be a consequence of $\mathbb{CB}_t\mathbf{R}_t$. Thus we need to modify the notion of common belief slightly when applied to the event "player i is rational at date t", so that the belief operator of player i at time t is never followed by the event stating that player i herself is rational at time t.

Recall that $\mathbf{R}_{i,t}$ is the event that player i is rational at time t (according to Definition 5) and that $B_{j,t}$ is the belief operator of player j at time t. The event that, at date t, there is common belief that player i is rational at date t, denoted by $\mathbf{CBR}_{i,t}$, is defined as follows:

$$\mathbf{CBR}_{i,t} = \left(\bigcap_{\substack{<j_1,...,j_m> \in N^* \\ j_m \neq i}} B_{j_1,t}...B_{j_m,t}\mathbf{R}_{i,t} \right). \tag{4}$$

Note that (4) rules out events of the form $B_{j_1,t}...B_{j_m,t}B_{i,t}\mathbf{R}_{i,t}$ for any - possibly empty - sequence of operators $B_{j_1,t}, ..., B_{j_m,t}$. Finally, let

$$\mathbf{CBR}_t = \bigcap_{i \in N} \mathbf{CBR}_{i,t}. \tag{5}$$

Remark 9. It follows from a standard argument that $\mathbf{CBR}_t \subseteq B_{i,t}\mathbf{CBR}_t,\ \forall i \in N, \forall t \in T$.[26]

Remark 10. In games where there are no simultaneous moves, $\mathbf{CBR}_t = \Omega$ and thus $\mathbf{R}_t \cap \mathbf{CBR}_t = \mathbf{R}_t$.[27]

[26] Proof. Suppose that $\omega \in \mathbf{CBR}_t$ but $\omega \notin B_{i,t}\mathbf{CBR}_t$. Then $\exists \omega' \in \mathcal{B}_{i,t}(\omega)$ such that $\omega' \notin \mathbf{CBR}_t$, that is, $\exists j \in N$ such that $\omega' \notin \mathbf{CBR}_{j,t}$. Then, by (4), $\exists < k_1, \ldots, k_m >\in N^*$ ($m \geq 1, k_m \neq j$) such that $\omega' \notin B_{k_1,t} \ldots B_{k_m,t}\mathbf{R}_{j,t}$ and, therefore, since $\omega' \in \mathcal{B}_{i,t}(\omega)$, $\omega \notin B_{i,t}B_{k_1,t} \ldots B_{k_m,t}\mathbf{R}_{j,t}$ and thus $\omega \notin \mathbf{CBR}_{j,t}$, which implies, since $\mathbf{CBR}_t \subseteq \mathbf{CBR}_{j,t}$ that $\omega \notin \mathbf{CBR}_t$, yielding a contradiction.

[27] This follows from the fact that events of the form $B_{i,t}\mathbf{R}_{i,t}$ are excluded from (4) and the fact that inactive players believe everything (see Definition 3), so that $\mathbf{CBR}_{i,t} = \Omega, \forall i \in N, \forall t \in T$.

The definition of forward belief of rationality remains essentially the same as in Section 3, the only exceptions being that (1) the events $\mathbf{R}_t$ are now replaced by $(\mathbf{R}_t \cap \mathbf{CBR}_t)$ and (2) the operators B_t are now replaced by $\mathbb{CB}_t$. Thus the event **FBR** is defined as follows. First, for $0 \leq k \leq m-1$ define $\mathbf{FBR}_k$ recursively by: $\mathbf{FBR}_{m-1} = \mathbf{R}_{m-1} \cap \mathbf{CBR}_{m-1}$, and, for $k < m-1$, $\mathbf{FBR}_k = (\mathbf{R}_k \cap \mathbf{CBR}_k) \cap \mathbb{CB}_k \mathbf{FBR}_{k+1} \cap \mathbf{FBR}_{k+1}$. Finally, define $\mathbf{FBR} = \mathbf{FBR}_0$.
The interpretation changes accordingly: **FBR** now refers not only to future but also to present rationality. Thus the word 'future' should be replaced with the expression 'present and future'. Since in games without simultaneous moves $(\mathbf{R}_t \cap \mathbf{CBR}_t) = \mathbf{R}_t$ (Remark 10) and $\mathbb{CB}_t E = B_t E$ for every event E and date t (Remark 8), in games without simultaneous moves the modified definition of **FBR** coincides with the definition given in Section 3.

4.4. Generalized backward induction procedure

The generalized backward induction procedure defined in Section 3 needs to be modified as follows: at every stage of the procedure instead of merely eliminating strictly dominated choices we perform an iterative elimination of strictly dominated choices at the information set(s) under consideration. The details are spelled out in the Appendix. As an example, consider the game of Figure 1. The algorithm starts at decision history a and iteratively deletes first choice g for Player 2 (strictly dominated by f), then choice d for Player 1 (which has now become strictly dominated by c) and finally choice f for Player 2, thus leaving only the profile (c, e). At the second stage a is deleted for Player 1 at the root (since it is now strictly dominated by b). Thus the algorithm yields a unique terminal history, namely b. By Proposition 1 it follows that, in any model of this game, if ω is a state such that $\omega \in \mathbf{FBR}$ then $\zeta(\omega) = b$.

5. Discussion

In this section we discuss several aspects of the proposed framework and related literature.

5.1. Models and predictions

As noted in the introduction, the standard models used in the literature are static structures where players are modeled as choosing strategies 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 will be reached, but if it were to happen then I would have such and such beliefs") and (3) interpersonal hierarchical constructions involving them ("I believe that if I were to play a then Player 2 would be surprised, would play b and would form the belief that I would subsequently erroneously believe that she had 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: a state 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 (and would have believed) 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 policy is postulated or necessary (for a discussion of this point see Section 5.3

below). 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 [19]). These conditionals are meant to capture the "exploratory" beliefs of the active player(s) ("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 she is about to do (see Remark 2 above and the discussion of the philosophical literature on this point contained in Section 4 of [11]).

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 doxastic hypothesis is a *set of outcomes or terminal histories*, not a set of strategy profiles. 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 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. For example, 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 [2] (Theorem 6.1, p. 244) 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.[28] It is not clear why 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 assessment, or prediction, as better or more reasonable than another. For example, [2] shows that in the game of Figure 1 above his notion of common belief in future rationality and the notion of extensive-form rationalizability ([21, 22, 23, 6]) yield the same prediction in terms of outcome (namely terminal history b, which is also the prediction of our notion of forward belief of rationality) but different counterfactual predictions at the unreached history a. Do we really need to settle those counterfactuals?

5.2. Comparison with [2]

As noted in the introduction, the content of this paper is closely related to the ideas put forward in [1, 2]. We already pointed out the most important difference between Perea's notion of common belief of future rationality (CBFR) and our notion of forward belief of rationality (FBR), namely that the former is expressed in terms of static hierarchies of beliefs about strategies, while the latter is expressed in terms of temporal beliefs about choices. Here we point out further differences.

[28] [17] also derives the entire backward-induction strategy profile from the hypothesis of common knowledge of rationality [as noted in [20, Footnote 4, p. 194], Aumann *proves* that common knowledge of substantive rationality implies the backward-induction *strategies* but *states* the weaker claim that it implies the backward-induction *path*].

Our generalized backward induction (GBI) algorithm is conceptually very similar to the backward dominance (BD) procedure proposed by [2].[29] 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" ([2, p. 244]).[30] 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 is *a set of terminal histories*.[31] The BD procedure yields the strategies that can rationally be chosen under CBFR if one does not impose common belief in Bayesian updating; our GBI procedure yields the set of *outcomes* that are compatible with the notion of FBR. It should be noted that the notion of CBFR is stronger than that of FBR. First of all, the former imposes the requirement that a player, at every information set, believes that all opponents, at *all past and future* information sets, believe in their opponents' future rationality, while the latter only requires that a player, at every information set, believes that all *future* opponents, at all future information sets, believe in their opponents' future rationality. A more important difference is a consequence of the fact that CBFR imposes restrictions on *strategies* while FBR imposes restrictions only on *choices*. To see why this is important, consider the game shown in Figure 5, known as the "battle of the sexes with an outside option".[32] The notion of FBR does not rule out any outcome in this game, in particular both *Ibc* and *Ibd* are compatible with FBR. This can be seen from the model illustrated in Figure 5, where $\mathbf{FBR} = \{\gamma, \epsilon, \theta\}$.[33] Thus, for example, ϵ is a state where there is forward belief in rationality and the outcome is *Ibc*. One could argue that a rational Player 1 would not plan to first choose *I* and then *b*, since this plan would yield her at most a utility of 1, whereas she can guarantee herself a utility of 2 by choosing *O*. Any strategy-based approach would rule out strategy (I, b) for Player 1 in this game, as it can never be optimal at time 0 for any belief that Player 1 can hold about Player 2's future choice. Indeed CBFR rules out strategy profiles that yield outcomes *Ibc* and *Ibd*. The fact that *Ibc* and *Ibd* are compatible with FBR can be viewed as a weakness of the notion of FBR. However, it is consistent with a strictly behavioral point of view according to which, at date 0, Player 1 can only *choose* between *I* and *O* and cannot commit to a future choice between *a* and *b*: she can have beliefs about what she will choose at date 1, but she cannot make that choice at time 0. At state ϵ in the model of Figure 5 Player 1 is rational in choosing *I* because she believes that afterwards she herself will choose *a* and Player 2 will follow with *c*; however, she is mistaken in her belief about her future action, because at state ϵ she will actually choose *b* (and rationally so, since at date 1 her belief about Player 2's subsequent choice has changed from *c* to *d*). One could make it a part of the definition of rationality that a player always correctly

[29] Similar procedures are the "backwards rationalizability procedure" of [1] and the "iterated conditional dominance procedure" of [24] (see also [25]), which selects the strategies that correspond to the notion of extensive-form rationalizability ([21, 22, 23, 6]). For a detailed discussion of how they relate to each other see [2, Section 7].

[30] On the other hand, the *backward rationalizability* procedure of [1] 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.

[31] 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.

[32] I am grateful to an anonymous reviewer for suggesting a discussion of this example.

[33] $\mathbf{R}_0 = \Omega\backslash\{\beta\}, \mathbf{R}_1 = \Omega\backslash\{\delta, \eta\}, \mathbf{R}_2 = \Omega\backslash\{\alpha, \lambda\}, B_0R_1 = B_0R_2 = B_1R_2 = B_0B_1R_2 = \Omega$ and thus $\mathbf{FBR} = R_0 \cap R_1 \cap R_2 = \{\gamma, \epsilon, \theta\}$.

anticipates (that is, has no uncertainty and has correct beliefs about) her own choices at later dates. Such a requirement would reflect the view that for a player "choosing" a plan of action (or strategy) means nothing more than forming a precise and correct belief about her own future actions.[34] Indeed, if one adds the requirement that a player correctly anticipate her own *actual* future choices,[35] then, in the game of Figure 5, outcomes *Ibc* and *Ibd* become incompatible with forward belief of rationality.

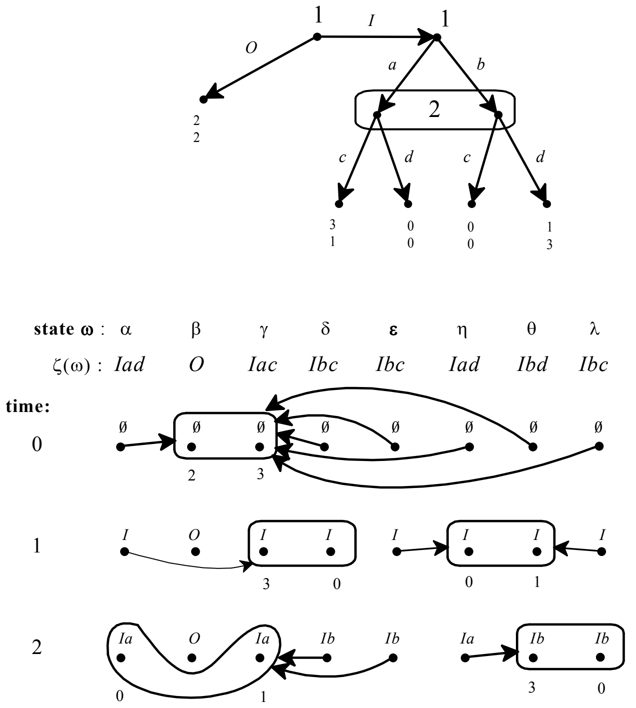


Figure 5: The battle of the sexes with outside option and a model of it.

5.3. Belief revision and subjective counterfactuals

In this section we clarify the claim made in the introduction that in the framework proposed in this paper belief revision and counterfactuals play no role.

[34] The interpretation of the notion of strategy as a belief concerning one's own future behavior can be found, for example, in [18, p. 391]: "A more intuitive interpretation of a strategy - in particular, the correct one in contexts where players cannot delegate their choices to mechanical devices - is, however, that of a subjective plan describing what the player thinks she would choose, at each history of hers, should that history occur. Under this interpretation, prescriptions of actions to be chosen at unreached histories can only reflect a player's belief about her own behavior, conditional on counterfactual events."

[35] Formally: if $t' > t$, $i \in \zeta_t(\omega) \cap \zeta_{t'}(\omega)$ and a_i is the action that player i takes at $\zeta_{t'}(\omega)$ then, for every $\omega' \in \mathcal{B}_{i,t}(\omega)$, if $\zeta_{t'}(\omega')$ belongs to the same information set as $\zeta_{t'}(\omega)$ then the action that player i takes at $\zeta_{t'}(\omega')$ is a_i.

There are two possible ways of understanding the expression "belief revision". The simplest interpretation of "belief revision" is "a change in belief prompted by new information". We call this *factual belief revision*. In this sense belief revision is, of course, a necessary aspect of the temporal beliefs that characterize our framework. For example, at date 0 Player 1 might believe that she will not have to make any further choices at future dates, but then discover that she was wrong. For instance, one can construct a model of the extensive form shown in Figure 6 where at some state, say α, the actual play is *bce* and yet, at date 0, Player 1 erroneously believes that if she chooses b then Player 2 will follow with d and then g, after which Player 3 will play m. Since at state α Player 1 plays b and subsequently Player 2 actually chooses c, at date 2 Player 1 unexpectedly finds herself having to choose again (between e and f). Hence such a model would necessarily incorporate a belief change, or factual belief revision, for Player 1. Note, however, that the model does *not* specify a belief of Player 1 at date 0 concerning what she would do and believe at date 2 if - contrary to her expectations - Player 2 played c.

However, the typical interpretation of "belief revision" is in terms of a complex epistemic state that incorporates initial beliefs as well as a set of conditional beliefs that encode the player's disposition to change her beliefs in response to all possible and relevant items of information, that is, "belief revision" means a "complete belief revision policy"; we call this *dispositional belief revision*. Belief revision in this sense is also referred to in the literature as "subjective counterfactuals" and is a *necessary ingredient of static analyses of dynamic games*, since a player needs to reason about her own and others' behavior even at those information sets that are ruled out by her initial beliefs.[36]

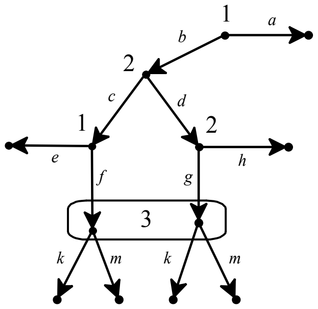


Figure 6: An extensive form used to discuss belief revision.

Belief revision in this sense is definitely *not* an ingredient of the class of models considered in this paper, where the only beliefs that are specified are those of the active players at the time of choice and concern past moves and future possible developments in the game. For example, the belief of Player 1 described above for the extensive form of Figure 6 (the erroneous belief, at state α and time 0, that if she plays b then the outcome will be *bdgm*, while the actual play at α is *bce*) does not incorporate a conditional belief of the form "what would I do and what would I believe about Player 3's choice if I discovered that Player 2 played c?" There will be such a

[36] Subjective counterfactuals are typically modeled either by means of conditional probability systems (see, for example, [6]) or by means of subjective selection functions (see, for example, [26, 12]), or similar structures. For a detailed discussion see [16].

factual belief at state α and time 2 (and it might very well involve a belief about Player 3's future choice that differs from the earlier one at date 0) and thus there will be "factual belief revision", but the more complex epistemic state encoded in dispositional belief revision is not postulated and is not needed.

Furthermore, it is also clear that objective (sometimes called "causal") counterfactuals are completely absent from our framework, which is a purely subjective, doxastic framework. To continue our example based on the extensive form of Figure 6, at state α, where Player 2 actually plays c, there is no way of evaluating the "objective" counterfactual "if Player 2 had played d and then g then Player 3 would have played m." One can enrich the framework by adding a Stalnaker-Lewis "selection function" ([27, 28]) thereby providing the necessary tools for evaluating such counterfactuals (and beliefs about them), but it is not clear what one would gain from the added complexity.

We conclude this discussion by stressing that we do not claim that the proposed framework completely avoids conditionals or hypothetical thinking. What we do claim is that the analysis is developed without the tools that are identified with the notion of "counterfactual" in the philosophy literature, namely the Stalnaker-Lewis objective selection functions or their subjective counterparts which capture dispositional belief revision (or conceptually equivalent structures, such as conditional probability systems). The conditionals involved in the class of models considered here are much simpler: for instance, a player only forms beliefs about the choices that her opponents (and she herself) will make at information sets she *actually* deems possible.

5.4. Backward induction versus forward induction

Like Perea's notion of CBFR, the notion of forward belief of rationality captures a general concept of backward induction. In the literature a contrasting notion has been discussed, namely the notion of forward induction (see, for example, [22, 23, 6, 29, 21]). As noted in [30], one cannot find a precise definition of forward induction, but the two notions represent different types of reasoning:

> "Backward induction represents a pattern of reasoning in which a player, at every stage of the game, only reasons about the opponents' future behavior and beliefs, and not about choices that have been made in the past. So, he takes the opponents' past choices for granted, but does not draw any new conclusions from these.
>
> In contrast, forward induction requires a player, at every stage, to think critically about the observed past choices by his opponents. He should always try to find a plausible reason for why his opponents have made precisely these choices in the past, and he should use this to possibly reconsider his belief about the opponents' future, present, and unobserved past choices" [30, p. 169].

The two notions often yield different results: for example, in the game of Figure 1 Perea's notion of CBFR uniquely selects strategy e for Player 2, while forward induction (as captured by the notion of extensive-form rationalizability) singles out strategy f for Player 2 (although the predicted outcome is the same, namely b).

The notion of forward induction has so far been modeled in the standard static approach in terms of strategies and conditional beliefs. Can the idea of forward induction be captured in the class of models considered in this paper? There are two aspects of the notion of forward induction as developed so far:

1. The first is a “factual” aspect of beliefs: players’ beliefs reflect, if possible, the conviction that earlier movers acted rationally.
2. The second is a restriction on players’ belief revision policy: when faced with information that contradicts earlier beliefs, players should (if possible) switch to a new belief that does not question the rationality of earlier movers.[37]

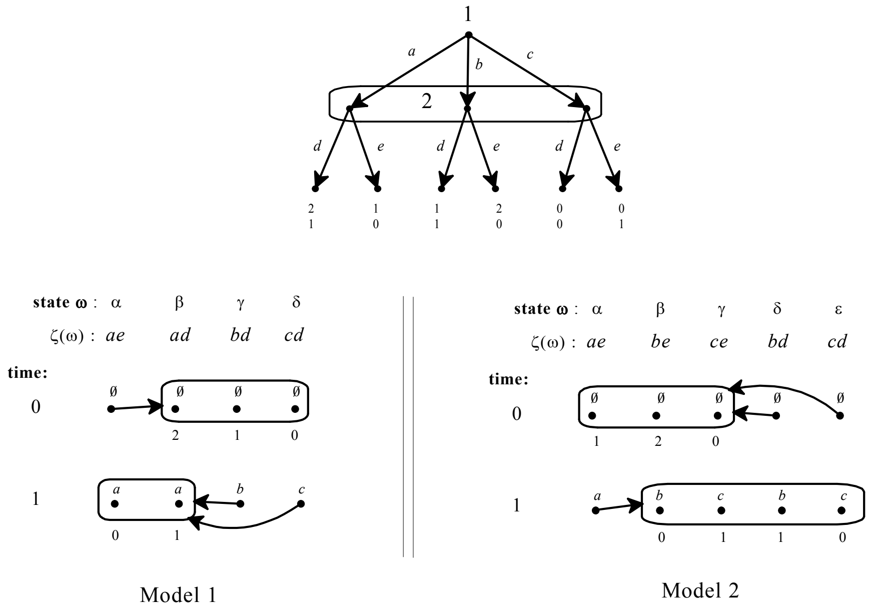
The first requirement can easily be incorporated in the class of models considered here, by adding to the belief that future players will be rational the belief that past players acted rationally. For example, consider the game of Figure 7 (which reproduces Figure 3 of [2]). For this game Perea’s notion of CBFR and our notion of FBR yield the same prediction, namely the set of outcomes $\{ad, ae, bd, be\}$ (CBFR yields strategies a and b for Player 1 and d and e for Player 2). Since choice c is strictly dominated, if Player 1 is rational at date 0, she will not choose c; hence if Player 2 believes, at date 1, that Player 1 chose rationally at date 0 ($B_1\mathbf{R}_0$), then he believes that Player 1 did not choose c. Hence, if he is rational at date 1, he will not choose e; thus if Player 1 believes, at date 0, that Player 1 will believe, at date 1, that Player 1 was rational at date 0 ($B_0B_1\mathbf{R}_0$) and, furthermore, Player 1 believes that Player 2 will be rational at date 1, then she will play a and the outcome will be ad. In Figure 7 two models of the game are shown. For this game the event $\mathbf{FBR}$ is given by $\mathbf{FBR} = \mathbf{R}_0 \cap \mathbf{R}_1 \cap B_0\mathbf{R}_1$. In Model 1 we have that $\mathbf{R}_0 = \{\alpha, \beta\}, \mathbf{R}_1 = \{\beta, \gamma, \delta\}$, $B_0\mathbf{R}_1 = \Omega$ and thus $\mathbf{FBR} = \{\beta\}$; furthermore, $B_1\mathbf{R}_0 = B_0B_1\mathbf{R}_0 = \Omega$ and thus $\mathbf{FBR} \cap\ B_1\mathbf{R}_0 \cap B_0B_1\mathbf{R}_0 = \{\beta\}$ and indeed $\zeta(\beta) = ad$. On the other hand, in Model 2, $\mathbf{R}_0 = \{\beta, \delta\}$ and $\mathbf{R}_1 = B_0\mathbf{R}_1 = \Omega$ so that $\mathbf{FBR} = \{\beta, \delta\}$ (and thus it is consistent with forward belief of rationality for Player 1 to choose b); on the other hand, $B_1\mathbf{R}_0 = B_0B_1\mathbf{R}_0 = \varnothing$ and thus $\mathbf{FBR} \cap B_1\mathbf{R}_0 \cap B_0B_1\mathbf{R}_0 = \varnothing$.

Indeed it can be shown that for every model of the game of Figure 7, if ω is a state such that $\omega \in \mathbf{FBR} \cap B_1\mathbf{R}_0 \cap B_0B_1\mathbf{R}_0$ then $\zeta(\omega) = ad$, thus capturing some aspects of the notion of forward induction.

We now turn to the second view of forward induction, namely as a restriction on players’ belief revision policy: when faced with information that contradicts earlier beliefs, players should - if possible - switch to a new belief that does not question the rationality of earlier movers. As remarked above, the standard static approach needs to rely on subjective counterfactuals or dispositional belief revision and the added restriction imposed by the idea of forward induction poses a problem, namely that in a particular model there may not be “enough belief hierarchies” to rationalize an observed deviation by an opponent. Thus the literature on forward induction has been developed within a *complete* epistemic model that contains all possible belief hierarchies (e.g. the infinite universal space of [6]) or an otherwise “sufficiently rich” model (e.g. [7]). On the other hand, in our approach dispositional belief revision plays no role and the type of characterization results that can be obtained are statements about *the class of models*, namely statements of the form “*if* at a state ω of an arbitrary model of the game a certain condition holds, then at ω fact x holds”; there may be models where the condition does not hold at any state (e.g. Model 2 in Figure 7) and models where the condition is satisfied at some state (e.g. Model 1 in Figure 7; for the game of Figure 7 the condition we discussed was expressed by the event $\mathbf{FBR} \cap B_1\mathbf{R}_0 \cap B_0B_1\mathbf{R}_0$).

The question of how to adequately formulate and capture the notion of forward induction in this class of models is an open one and certainly worth pursuing. It is, however, beyond the scope

[37] For example, [22, p. 179] writes “A player should always try to interpret her information about the behavior of her opponents assuming that they are not implementing ‘irrational’ strategies.”



Model 1

Model 2

Figure 7: Forward and backward belief in rationality.

of this paper.

We conclude by acknowledging our intellectual debt to Samet ([12]) who introduced behavior-based models, where states are described in terms of actual choices rather than strategies. His analysis relies on subjective counterfactuals, expressed in terms of hypothetical knowledge operators (which are a version of subjective selection functions: see [26] for a detailed discussion). Samet restricts attention to perfect-information games and provides a characterization of the backward-induction *outcome* in terms of a condition that he calls "common hypothesis of node rationality." 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" [12, p.233]), it assumes knowledge, rather than belief, it imposes the condition that if at a state a player considers it possible that she takes an action at a decision history then she knows that, if that decision history is reached, she takes that action and, finally - as noted above - he makes essential use of subjective counterfactuals.

Appendix A. Proofs

We prove Propositions 1 and 2 for the general case of von Neumann extensive games with possibly simultaneous moves. First we recall the definitions of forward belief of rationality and of the generalized backward induction procedure for the general case.

Recall (see Section 4.3) that when simultaneous moves are allowed the event **FBR** is defined as follows. First, for $0 \leq k \leq m-1$ define $\mathbf{FBR}_k$ recursively by: $\mathbf{FBR}_{m-1} = \mathbf{R}_{m-1} \cap \mathbf{CBR}_{m-1}$, and, for $k < m-1$, $\mathbf{FBR}_k = (\mathbf{R}_k \cap \mathbf{CBR}_k) \cap \mathbb{CB}_k \mathbf{FBR}_{k+1} \cap \mathbf{FBR}_{k+1}$. Finally, define $\mathbf{FBR} = \mathbf{FBR}_0$.

For the generalized backward induction procedure we start by defining the set $U(h)$ of *(iteratively) undominated choices* at decision history h: $U(h) = \prod_{i\in\iota(h)} (A_i(h)\backslash D_i(h))$, where $D_i(h)$ is the set of player i's strictly dominated choices at i's information set that contains h and is defined by: $D_i(h) = \bigcup_{j=0}^{\infty} D_i^j(h)$ and $D_i^j(h)$ is given recursively as follows (for $i \in \iota(h)$ and $j \geq 0$):

1. If $\ell(h) = \ell^{max} - 1$,
   - 1.a. $a_i \in D_i^0(h)$ if and only if $a_i \in A_i(h)$ and there exists a $b_i \in A_i(h)$ such that, for every $h' \in I_i(h)$, and for every $a_{-i} \in \prod_{k\in\iota(h')\backslash\{i\}} A_k(h')$, $u_i(h'(a_i, a_{-i})) < u_i(h'(b_i, a_{-i}))$ [that is, b_i yields higher utility than a_i at every history in the information set containing h, when combined with any profile of actions of the active players other than i (if any); in other words, if a_i is strictly dominated by some other choice b_i at the information set of player i that contains h].
   - 1.b. For $j > 0, a_i \in D_i^j(h)$ if and only if $a_i \in A_i(h)$ and there exists a $b_i \in A_i(h)$ such that, for every $h' \in I_i(h)$, and for every $a_{-i} \in \prod_{k\in\iota(h')\backslash\{i\}} \left(A_k(h')\backslash D_k^{j-1}(h')\right)$, $u_i(h'(a_i, a_{-i})) < u_i(h'(b_i, a_{-i}))$ [that is, a_i is strictly dominated by some other choice b_i if, at every history in the information set containing h, the other players are assumed to play only actions that have not (so far) been marked as dominated; note that, $\forall j \geq 0, D_i^{j+1}(h) \supseteq D_i^j(h)$].
2. Having defined $U(h)$ for every decision history h such that $\ell(h) = k$, with $0 < k \leq \ell^{max} - 1$, define, for h such that $\ell(h) = k - 1$, $U(h) = \prod_{i\in\iota(h)}(A_i(h)\backslash D_i(h))$, where $D_i(h) = \bigcup_{j=0}^{\infty} D_i^j(h)$ and $D_i^j(h)$ is given recursively as follows (for $i \in \iota(h)$ and $j \geq 0$):
   - 2.a. $a_i \in D_i^0(h)$ if and only if $a_i \in A_i(h)$ and there exists a $b_i \in A_i(h)$ such that, $\forall h' \in I_i(h), \forall a_{-i} \in \prod_{k\in\iota(h')\backslash\{i\}} A_k(h')$, $u_i(z) < u_i(z')$ for all $z, z' \in Z$ such that $z = h'(a_i, a_{-i})c_1 \dots c_p$ $(p \geq 0)$, $z' = h'(b_i, a_{-i})d_1 \dots d_q$ $(q \geq 0)$ with $c_1 \in U(h'(a_i, a_{-i}))$, $d_1 \in U(h'(b_i, a_{-i}))$ and, $\forall j = 2, \dots, p$, $\forall k = 2, \dots, q$, $c_j \in U(h'(a_i, a_{-i})c_1 \dots c_{j-1})$ and $d_k \in U(h'(b_i, a_{-i})d_1 \dots d_{k-1})$ [that is, a_i is strictly dominated by some b_i at the information set of player i that contains h, assuming that after the move at that information set only undominated choices are made].
   - 2.b. For $j > 0, a_i \in D_i^j(h)$ if and only if $a_i \in A_i(h)$ and there exists a $b_i \in A_i(h)$ such that, $\forall h' \in I_i(h), \forall a_{-i} \in \prod_{k\in\iota(h')\backslash\{i\}} \left(A_k(h')\backslash D_k^{j-1}(h')\right)$, $u_i(z) < u_i(z')$, for all $z, z' \in Z$ such that $z = h'(a_i, a_{-i})c_1 \dots c_p$ $(p \geq 0)$, $z' = h'(b_i, a_{-i})d_1 \dots d_q$ $(q \geq 0)$ with $c_1 \in U(h'(a_i, a_{-i}))$, $d_1 \in U(h'(b_i, a_{-i}))$ and, $\forall j = 2, \dots, p$, $\forall k = 2, \dots, q$, $c_j \in U(h'(a_i, a_{-i})c_1 \dots c_{j-1})$ and $d_k \in U(h'(b_i, a_{-i})d_1 \dots d_{k-1})$ [that is, a_i is strictly dominated by some b_i, assuming that at every history in the information set containing h the other players only take actions that have not (so far) been marked as dominated and, furthermore, after the move at that information set only undominated choices are made].

For example, in the game of Figure 1, $D_1^0(a) = \varnothing$, $D_2^0(a) = \{g\}$, $D_1^1(a) = \{d\}$, $D_2^1(a) = \{f, g\}$ and after this no other choices are strictly dominated, so that $D_1(a) = \{d\}$ and $D_2(a) = \{f, g\}$ and thus $U(a) = \{(c, e)\}$, so that $D_1(\emptyset) = \{a\}$ and $U(\emptyset) = \{b\}$.

The function $f_{BI} : H \to 2^Z$ and the event **BI** are given by: (1) if $h \in Z$ then $f_{BI}(h) = \{h\}$ and (2) if h is a decision history then $f_{BI}(h) = \{z \in Z : z = ha_1a_2...a_m,\ a_1 \in U(h)\ and\ a_j \in U(ha_1...a_{j-1}), \forall j = 2, ..., m\}$, that is, $f_{BI}(h)$ is the set of terminal histories that can be reached from h by following only undominated choices. Finally, $\mathbf{BI} = f_{BI}(\emptyset)$.

Proof of Proposition 1. Fix a von Neumann extensive game (with possibly simultaneous moves) and a model of it. Let $m = \ell^{max}$ be the depth of the game. First we prove the following:

$$\begin{array}{l}\forall \omega \in \Omega, \text{ if } \omega \in \mathbf{R}_{m-1} \cap \mathbf{CBR}_{m-1} \text{ then, } \forall i \in \iota(\zeta_{m-1}(\omega)), \text{ if } a_i \\ \text{is the action taken by } i \text{ at } \zeta_{m-1}(\omega), \text{ then } a_i \notin D_i(\zeta_{m-1}(\omega)).\end{array} \tag{A.1}$$

As a first step we show that

$$\begin{array}{l}\forall \omega \in \Omega, \forall i \in \iota(\zeta_{m-1}(\omega)), \text{ if } \omega \in \mathbf{R}_{i,m-1} \text{ and } a_i \text{ is the} \\ \text{action taken by } i \text{ at } \zeta_{m-1}(\omega), \text{ then } a_i \notin D_i^0(\zeta_{m-1}(\omega)).\end{array} \tag{A.2}$$

Let $\omega \in \Omega$ and $i \in \iota(\zeta_{m-1}(\omega))$ be such that $\omega \in \mathbf{R}_{i,m-1}$ and suppose that $a_i \in D_i^0(\zeta_{m-1}(\omega))$, where a_i is the action taken by player i at $\zeta_{m-1}(\omega)$. Then there exists a $b_i \in A_i(\zeta_{m-1}(\omega))$ such that

$$\forall h' \in I_i(\zeta_{m-1}(\omega)), \forall c_{-i} \in \prod_{k \in \iota(h') \setminus \{i\}} A_k(h'), \quad u_i(h'(b_i, c_{-i})) > u_i(h'(a_i, c_{-i})). \tag{A.3}$$

Fix an arbitrary $\omega' \in \mathcal{B}_{i,m-1}(\omega)$ (recall that, by 2.1 of Definition 3, $\mathcal{B}_{i,m-1}(\omega) \neq \varnothing$) and let $c_{-i} \in A_{-i}(\zeta_{m-1}(\omega'))$ be the action profile played by the players other than i at $\zeta_{m-1}(\omega')$ (if any). By 2.3s of Definition 6 there exist $\omega_1, \omega_2 \in \mathcal{B}_{i,m-1}(\omega)$ such that $\zeta_m(\omega_1) = \zeta_{m-1}(\omega')(a_i, c_{-i})$ and $\zeta_m(\omega_2) = \zeta_{m-1}(\omega')(b_i, c_{-i})$. Then, by (A.3), at state ω and time $m-1$ player i believes that b_i is better than a_i (see Footnote 25) and thus, since a_i is her choice at $\zeta_{m-1}(\omega)$, she is not rational at $(\omega, m-1)$, that is, $\omega \notin \mathbf{R}_{i,m-1}$, contradicting our hypothesis.

Next we show that:

$$\begin{array}{l}\forall \in \Omega, \forall i \in \iota(\zeta_{m-1}(\omega)), \text{ if } \omega \in \mathbf{R}_{i,m-1} \cap \mathbf{CBR}_{m-1} \text{ then, if } a_i \\ \text{is the action taken by } i \text{ at } \zeta_{m-1}(\omega), \text{ then } a_i \notin D_i^1(\zeta_{m-1}(\omega)).\end{array} \tag{A.4}$$

Let $\omega \in \Omega$ and $i \in \iota(\zeta_{m-1}(\omega))$ be such that $\omega \in \mathbf{R}_{i,m-1} \cap \mathbf{CBR}_{m-1}$. Fix an arbitrary $\omega' \in \mathcal{B}_{i,m-1}(\omega)$ and an arbitrary $j \in \iota(\zeta_{m-1}(\omega')) \setminus \{i\}$. Since $\mathbf{CBR}_{m-1} \subseteq \mathbf{CBR}_{j,m-1}$, $\omega \in \mathbf{CBR}_{j,m-1}$ and thus, by definition of $\mathbf{CBR}_{j,m-1}$ (see (5)) and the fact that $j \neq i$,

$$\omega \in B_{i,m-1}\mathbf{R}_{j,m-1}. \tag{A.5}$$

Hence $\mathcal{B}_{i,m-1}(\omega) \subseteq \mathbf{R}_{j,m-1}$ and thus $\omega' \in \mathbf{R}_{j,m-1}$; hence, by (A.2), if a_j is the action taken by j at $\zeta_{m-1}(\omega')$ then $a_j \notin D_j^0(\zeta_{m-1}(\omega'))$. Thus at $(\omega, m-1)$ player i believes that at every history that she considers possible the action profile taken there by the other players (if any) is not "level-0 strictly dominated". Hence, since player i is rational at $(\omega, m-1)$ ($\omega \in \mathbf{R}_{i,m-1}$), if a_i is the action taken by i at $\zeta_{m-1}(\omega)$, then $a_i \notin D_i^1(\zeta_{m-1}(\omega))$.
Next we prove by induction that:

$$\begin{array}{l}\forall \in \Omega, \forall i \in \iota(\zeta_{m-1}(\omega)), \text{ if } \omega \in \mathbf{R}_{i,m-1} \cap \mathbf{CBR}_{m-1} \text{ then, if } a_i \text{ is the} \\ \text{action taken by } i \text{ at } \zeta_{m-1}(\omega), \text{ then, } \forall k \geq 1,\ a_i \notin D_i^k(\zeta_{m-1}(\omega)).\end{array} \tag{A.6}$$

The base step, namely the case $k = 1$, is given by (A.4). Thus we only need to prove the induction step, that is, that if (A.6) true for $k \geq 1$ then it must be true for $k+1$. Let $\omega \in \Omega$ and

$i \in \iota\,(\zeta_{m-1}(\omega))$ be such that $\omega \in \mathbf{R}_{i,m-1} \cap \mathbf{CBR}_{m-1}$. Fix an arbitrary $j \in N\backslash\{i\}$. Then, as shown above, (A.5) must hold, that is, $\omega \in B_{i,m-1}\mathbf{R}_{j,m-1}$. By Remark 9, $\mathbf{CBR}_{m-1} \subseteq B_{i,m-1}\mathbf{CBR}_{m-1}$ and thus $\omega \in B_{i,m-1}\mathbf{CBR}_{m-1}$. Hence

$$\omega \in B_{i,m-1}\left(\mathbf{R}_{j,m-1} \cap \mathbf{CBR}_{m-1}\right). \tag{A.7}$$

Fix an arbitrary $\omega' \in \mathcal{B}_{i,m-1}(\omega)$. Then, by (A.7), $\omega' \in \mathbf{R}_{j,m-1} \cap \mathbf{CBR}_{m-1}$ and thus, by our induction hypothesis, if a_j is the action taken by player j at $\zeta_{m-1}(\omega')$ then $a_j \notin D_j^k(\zeta_{m-1}(\omega'))$. Thus at $(\omega, m-1)$ player i believes that at every history that she considers possible the action profile taken there by the other players (if any) is not "level-k strictly dominated". Hence, since $\omega \in \mathbf{R}_{i,m-1}$, if a_i is the action taken by i at $\zeta_{m-1}(\omega)$, then $a_i \notin D_i^{k+1}(\zeta_{m-1}(\omega))$. Thus (A.6) holds and hence, by definition of $D_i(\zeta_{m-1}(\omega))$, we have proved (A.1).

Next we prove that

$$\begin{array}{l}\text{For every } t \text{ with } 0 \le t \le m-1 \text{ and for every } \omega \in \Omega, \\ \text{if } \omega \in \mathbf{FBR}_t \text{ then } \zeta(\omega) \in f_{BI}(\zeta_t(\omega)).\end{array} \tag{A.8}$$

Fix arbitrary $t \in T$ and $\omega \in \mathbf{FBR}_t$. If $\zeta_t(\omega)$ is a terminal history, then $\zeta_t(\omega) = \zeta(\omega)$ (see Definition 2) and, by definition of $f_{BI}(\cdot)$, $f_{BI}(\zeta(\omega)) = \{\zeta(\omega)\}$. Thus $\zeta(\omega) \in f_{BI}(\zeta_t(\omega))$. The case where $\zeta_t(\omega)$ is a decision history is proved by induction.

Base step: $t = m-1$. Fix arbitrary $\omega \in \mathbf{FBR}_{m-1} = \mathbf{R}_{m-1} \cap \mathbf{CBR}_{m-1}$ and suppose that $\zeta_{m-1}(\omega)$ is a decision history. We need to show that if a is the action profile played at $\zeta_{m-1}(\omega)$ (that is, $\zeta_m(\omega) = \zeta_{m-1}(\omega)a$) then $a \in U\,(\zeta_{m-1}(\omega))$. But this is precisely what (A.1) states.

Induction step. Suppose that (A.8) is true for $t = k$ with $1 < k \le m-1$. We want to show that it is true for $t = k-1$. Fix an arbitrary state β and suppose that

$$\beta \in \mathbf{FBR}_{k-1} = (\mathbf{R}_{k-1} \cap \mathbf{CBR}_{k-1}) \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k \cap \mathbf{FBR}_k. \tag{A.9}$$

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. Fix arbitrary $i \in \iota(\zeta_{k-1}(\beta))$. By definition of common belief (see (3) and Remark 7), $\mathbb{CB}_{k-1}\mathbf{FBR}_k \subseteq B_{i,k-1}\mathbf{FBR}_k \cap B_{i,k-1}\mathbb{CB}_{k-1}\mathbf{FBR}_k = B_{i,k-1}(\mathbf{FBR}_k \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k)$. Thus, since, by (A.9), $\beta \in \mathbb{CB}_{k-1}\mathbf{FBR}_k$ it follows that $\beta \in B_{i,k-1}(\mathbf{FBR}_k \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k)$, so that, by the induction hypothesis, at state β and time $k-1$ player i believes (recall that, by Definition 3, $\mathcal{B}_{i,k-1}(\beta) \neq \varnothing$) that after the action profile taken at every history that he considers possible only terminal histories selected by the function $f_{BI}(\cdot)$ can be reached (that is, only terminal histories that are reached by undominated choices) and, furthermore, that this fact is common belief among the active players at time $k-1$. If we establish that if a_i is the action played by player i at $\zeta_{k-1}(\beta)$ then $a_i \notin D_i(\zeta_{k-1}(\beta))$ then, since player i was chosen arbitrarily from the set $\iota(\zeta_{k-1}(\beta))$, it follows that if a is the action profile played at $\zeta_{k-1}(\beta)$ then $a \in U(\zeta_{k-1}(\beta))$. Then to complete the proof it is sufficient to note that, by (A.9), $\beta \in \mathbf{FBR}_k$, so that, by the induction hypothesis, $\zeta(\beta) \in f_{BI}(\zeta_k(\beta))$; by Remark 6 it follows from this and the fact that $a \in U(\zeta_{k-1}(\beta))$, that $\zeta(\beta) \in f_{BI}(\zeta_{k-1}(\beta))$.

The proof that if a is the action profile played at $\zeta_{k-1}(\beta)$ then $a \in U(\zeta_{k-1}(\beta))$ is a repetition of the induction argument used to prove (A.1) and thus we will only highlight the key observations. First one shows the following (which corresponds to (A.2)):

$$\begin{array}{l}\forall \in \Omega, \forall j \in \iota\,(\zeta_{m-1}(\omega)), \text{ if } \omega \in \mathbf{R}_{j,k-1} \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k \text{ then, if } a_j \text{ is} \\ \text{the action taken by } j \text{ at } \zeta_{k-1}(\omega), \text{ then } a_j \notin D_j^0(\zeta_{k-1}(\omega)).\end{array} \tag{A.10}$$

This is done by using the fact that $\mathbb{CB}_{k-1}\mathbf{FBR}_k \subseteq B_{j,k-1}\mathbf{FBR}_k$ and thus at $(\omega, k-1)$ player j believes that after the profiles of actions taken at every decision history that she considers possible only undominated choices follow and thus, since $\omega \in \mathbf{R}_{j,k-1}$, $a_j \notin D_j^0(\zeta_{k-1}(\omega))$.

Then one shows the following (which corresponds to (A.4)):

$$\begin{array}{l}\forall \in \Omega,\ \forall i \in \iota\,(\zeta_{m-1}(\omega)),\ \text{if } \omega \in \mathbf{R}_{i,k-1} \cap \mathbf{CBR}_{k-1} \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k \\ \text{then, if } a_i \text{ is the action taken by } i \text{ at } \zeta_{k-1}(\omega), \text{ then } a_i \notin D_i^1(\zeta_{k-1}(\omega)).\end{array} \tag{A.11}$$

This is done by inferring from $\omega \in \mathbf{CBR}_{k-1} \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k$ that, for every $j \neq i$, $\omega \in B_{i,k-1}\left(\mathbf{R}_{j,k-1} \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k\right)$ and using (A.10) to conclude that, at every $\omega' \in \mathcal{B}_{i,k-1}(\omega)$, the active players at $\zeta_{k-1}(\omega')$ don't choose "level-0 strictly dominated actions", that is, at ω player i believes this fact and thus, since $\omega \in \mathbf{R}_{i,k-1}$, if a_i is the action taken by i at $\zeta_{k-1}(\omega)$, then $a_i \notin D_i^1(\zeta_{k-1}(\omega))$.

Now (A.11) becomes the base step for the proof by induction of the following claim (which corresponds to (A.6)):

$$\begin{array}{l}\forall \in \Omega,\ \forall i \in \iota\,(\zeta_{m-1}(\omega)),\ \text{if } \omega \in \mathbf{R}_{i,k-1} \cap \mathbf{CBR}_{k-1} \cap \mathbb{CB}_{k-1}\mathbf{FBR}_k \text{ then,} \\ \text{if } a_i \text{ is the action taken by } i \text{ at } \zeta_{k-1}(\omega), \text{ then, } \forall m \geq 1,\ a_i \notin D_i^m(\zeta_{k-1}(\omega)).\end{array} \tag{A.12}$$

As in the case of (A.6), the proof is a straightforward adaptation of the argument used to establish (A.11). ■

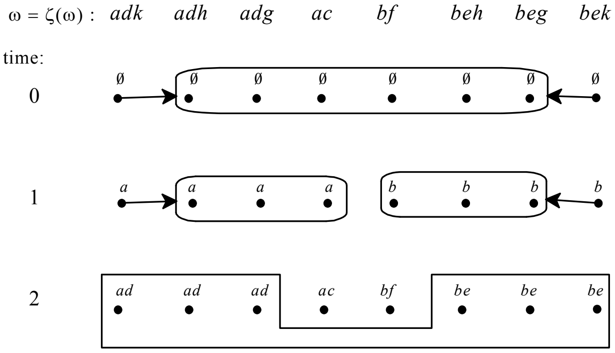


Figure A.8: The model described in the proof of Proposition 2 for the game of Figure 2.

Proof of Proposition 2. Fix a von Neumann extensive-form game and define 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 and $\iota(z_t) = \{i\}$ 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) $z'_t \in I_i(z_t)$ and (2) $z' \in f_{BI}(z'_{t+1})$. Figure A.8 shows the model just described for the game of Figure 2. If z_t is a decision history and the cardinality of $\iota(z_t)$ is greater than 1, then set $\mathcal{B}_{j,t}(z) = \varnothing$ for every player $j \notin \iota(z_t)$ and for each $i \in \iota(z_t)$ define $\mathcal{B}_{i,t}(z)$ as follows: $z' \in \mathcal{B}_{i,t}(z)$ if and only if (1) $z'_t \in I_i(z_t)$, (2) $z' \in f_{BI}(z'_{t+1})$ and (3) $\forall j \in \iota(z'_t)\backslash\{i\}$ if a_j is the action taken by player j at z'_t then $a_j \notin D_j(z'_t)$. Figure A.9 shows the model just described for the game of Figure 1. 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_i is the action taken by player i at z_t then $a_i \notin D_i(z_t)$. 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 $z \in \mathbf{FBR}_0 = \mathbf{FBR}$. ■

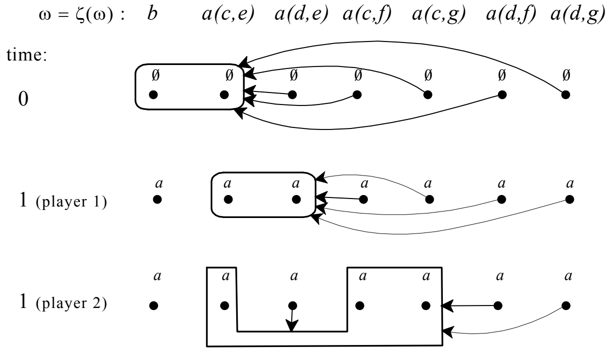


Figure A.9: The model described in the proof of Proposition 2 for the game of Figure 1.

References

[1] A. Penta, Robust dynamic mechanism design, Tech. rep., University of Wisconsin, Madison (2009). URL `http://www.econ.wisc.edu/ apenta/DMD.pdf`

[2] A. Perea, Belief in the opponents' future rationality, Games and Economic Behavior 83 (2014) 231–254.

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

[4] A. Perea, Epistemic foundations for backward induction: an overview, in: J. van Benthem, D. Gabbay, B. Löwe (Eds.), Interactive logic. Proceedings of the 7th Augustus de Morgan Workshop, Vol. 1 of Texts in Logic and Games, Amsterdam University Press, 2007, pp. 159–193.

[5] A. Baltag, S. Smets, J. Zvesper, Keep 'hoping' for rationality: a solution to the backward induction paradox, Synthese 169 (2009) 301–333.

[6] P. Battigalli, M. Siniscalchi, Strong belief and forward induction reasoning, Journal of Economic Theory 106 (2002) 356–391.

[7] A. Perea, Epistemic game theory: reasoning and choice, Cambridge University Press, Cambridge, 2012.

[8] R. Stalnaker, Belief revision in games: forward and backward induction, Mathematical Social Sciences 36 (1998) 31–56.

[9] R. Stalnaker, Extensive and strategic forms: games and models for games, Research in Economics 53 (1999) 293 –319.

[10] B. Skyrms, G. D. Bell, P. Woodruff, Theories of counterfactual and subjunctive conditionals in contexts of strategic interaction, Research in Economics 53 (1999) 275–291.

[11] G. Bonanno, A dynamic epistemic characterization of backward induction without counterfactuals, Games and Economics Behavior 78 (2013) 31–45.

[12] D. Samet, Hypothetical knowledge and games with perfect information, Games and Economic Behavior 17 (1996) 230–251.

[13] R. Aumann, On the centipede game, Games and Economic Behavior 23 (1998) 97–105.

[14] M. Osborne, A. Rubinstein, A course in game theory, MIT Press, Cambridge, 1994.

[15] P. Battigalli, G. Bonanno, Recent results on belief, knowledge and the epistemic foundations of game theory, Research in Economics 53 (1999) 149–225.

[16] G. Bonanno, Reasoning about strategies and rational play in dynamic games, in: J. van Benthem, S. Ghosh, R. Verbrugge (Eds.), Modeling strategic reasoning, Texts in Logic and Games, Springer, forthcoming.

[17] R. Aumann, Backward induction and common knowledge of rationality, Games and Economic Behavior 8 (1995) 6–19.

[18] P. Battigalli, A. Di-Tillio, D. Samet, Strategies and interactive beliefs in dynamic games, in: D. Acemoglu, M. Arellano, E. Dekel (Eds.), Advances in Economics and Econometrics. Theory and Applications: Tenth World Congress, Cambridge University Press, Cambridge, 2013, pp. 391–422.

[19] K. DeRose, The conditionals of deliberation, Mind 119 (2010) 1–42.

[20] D. Samet, Common belief of rationality in games of perfect information, Games and Economic Behavior 79 (2013) 192–200.

[21] D. Pearce, Rationalizable strategic behavior and the problem of perfection, Econometrica 52 (1984) 1029–1050.

[22] P. Battigalli, Strategic rationality orderings and the best rationalization principle, Games and Economic Behavior 13 (1996) 178–200.

[23] P. Battigalli, On rationalizability in extensive games, Journal of Economic Theory 74 (1997) 40–61.

[24] M. Shimoji, J. Watson, Conditional dominance, rationalizability, and game forms, Journal of Economic Theory 83 (1998) 161–195.

[25] J. Chen, S. Micali, The order independence of iterated dominance in extensive games, Theoretical Economics 8 (2013) 125–163.

[26] J. Halpern, Hypothetical knowledge and counterfactual reasoning, International Journal of Game Theory 28 (1999) 315–330.

[27] D. Lewis, Counterfactuals, Harvard University Press, 1973.

[28] R. Stalnaker, A theory of conditionals, in: N. Rescher (Ed.), Studies in logical theory, Blackwell, 1968, pp. 98–112.

[29] E. Kohlberg, J.-F. Mertens, On the strategic stability of equilibria, Econometrica 54 (1986) 1003–1038.

[30] A. Perea, Backward induction versus forward induction reasoning, Games 1 (2010) 168–188.