# MODAL LOGIC AND GAME THEORY: TWO ALTERNATIVE APPROACHES

Giacomo Bonanno*

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

e-mail: gfbonanno@ucdavis.edu

Keywords: modal logic, rationality, recommendation, internal consistency, backward induction.

**Abstract**

Two views of game theory are discussed: (1) game theory as a *description* of the behavior of *rational* individuals who recognize each other's rationality and reasoning abilities, and (2) game theory as an internally consistent *recommendation* to individuals on how to act in interactive situations. It is shown that the same mathematical tool, namely modal logic, can be used to explicitly model both views.

## 1. Introduction

Game theory can be thought of as being composed of two separate modules. The first module consists of a formal language for the description of interactive situations, that is, situations where several individuals take actions that affect each

[*] I am grateful for the comments of two anonymous referees, Pierpaolo Battigalli and the participants in the conference on Strategic Rationality in Economics (Sant'Arcangelo di Romagna, August 1999), where a first draft of this paper was presented.

other. This language provides alternative descriptions, from the more detailed one of extensive forms to the more condensed notions of strategic form and coalitional form. The language of game theory has proved to be useful in such diverse fields as economics, political science, military science, evolutionary biology, computer science, mathematical logic, experimental psychology, sociology and social philosophy. The unifying role of the game-theoretic language has been a major achievement in itself.

The second module is represented by the collection of *solution concepts*. Each solution concept associates with every game in a given class an outcome or set of outcomes. Most of the debate in game theory has centered on this module, in particular on the rationale for, and interpretation of, various solution concepts. From a broader point of view, the issue of debate is what the role and aims of game theory are (or should be). In this respect one can distinguish at least four different views of game theory:

**1.** Game theory as a *description* of how *rational* individuals behave:

> "Briefly put, game and economic theory are concerned with the interactive behavior of *Homo rationalis* - rational man. *Homo rationalis* is the species that always acts both purposefully and logically, has well-defined goals, is motivated solely by the desire to approach these goals as closely as possible, and has the calculating ability required to do so". (Aumann, 1985, p. 35)

**2.** Game theory as a *prescription* or *advice* to players on how to act.

**3.** Experimental game theory, whose aim is to *describe* the *actual* behavior of individuals.[1]

**4.** Evolutionary game theory, where outcomes are explained in terms of dynamic processes of natural selection.

In this paper we focus on the first two views of game theory and argue that the same tool, namely modal logic, can be used to model explicitly both of them.

The recent literature on the logical foundations of game-theoretic solution concepts has been concerned with the first view, namely game theory as a description

[1] In other words, experimental game theory tries to elucidate how *Homo sapiens* behaves (rather than *Homo rationalis*, to the extent that there is a difference between the two).

of the interactive behavior of *Homo rationalis*.[2] The relevance of modal logic is apparent from the fact that most papers in this literature make use, often only implicitly, of epistemic logic (that is, the logic of knowledge and belief) and try to determine what assumptions on the beliefs and reasoning of the players are implicit in various solution concepts. The task of this research program is to identify for any game the strategies that might be chosen by rational and intelligent players who know the structure of the game and the preferences of their opponents and who recognize each other's rationality and reasoning abilities. The main issue has been the relationship between the notion of common belief in rationality and non-cooperative solution concepts such as rationalizability and Nash equilibrium.

The second view of game theory, namely as a theory that *advises* players on how to behave in interactive situations, has been much less investigated. The objective here is not to capture the reasoning of rational players but rather to determine "what recommendations will be accepted, or followed, by free rational individuals" (Greenberg, 1990, p. 2). To put it differently, what consistency properties should be satisfied by a theory that purports to capture in its recommendations the players' incentives and goals?

It has long been recognized that solution concepts, such as Nash equilibrium, can be interpreted not as the outcome of players' reasoning but rather as game theory's recommendation:

> "The modern game-theoretical interpretation of equilibrium points in the sense of Nash (1951) ... is based on the idea that a rational theory should not be a self-destroying prophecy which creates an incentive to deviate for those who believe in it". (Selten, 1985, p. 79)

The notion that a recommendation should be consistent in the sense that it should not be a "self-destroying prophecy" was first introduced within cooperative game theory by von Neumann and Morgenstern (1947) and subsequently applied by Joseph Greenberg (1990) in his unifying theory of social situation. We will show in Section 4 that modal logic can be used to model explicitly also this second view of game-theoretic solution concepts.

The paper is organized as follows. In Section 2 we give a brief overview of modal logic. In Section 3 we discuss the epistemic interpretation of modal logic and give a flavor of the type of results proved within view (1) of game theory. In Section

[2] Extensive surveys of this literature are given in Battigalli and Bonanno (1999) and Dekel and Gul (1997).

4 we turn to a different interpretation of modal logic and use it to characterize one of the most frequently used solution concepts − namely backward induction in perfect information games − as a recommendation, thus reflecting view (2) of game theory. Section 5 concludes with a discussion of the advantages of using (modal) logic in the analysis of games.

## 2. Brief review of modal logic

Modal logic has two components: semantic and syntactic. The semantic component consists of a *frame* $\mathcal{F} = \langle \Omega, R_1, ..., R_m \rangle$, where $\Omega$ is a set of *states* (or possible worlds) and each $R_j$ $(j = 1, ..., m)$ is a binary relation on $\Omega$. If $\alpha R_j \beta$ we say that state $\beta$ is *$R_j$-accessible* from state $\alpha$. Depending on the context, the accessibility relations may be required to satisfy one or more properties, such as the following (for simplicity we drop the subscript of $R$):

- Seriality: $\forall \alpha \in \Omega, \exists \beta \in \Omega : \alpha R \beta$.
- Reflexivity: $\forall \alpha \in \Omega, \alpha R \alpha$.
- Transitivity: if $\alpha R \beta$ and $\beta R \gamma$ then $\alpha R \gamma$.
- Euclideanness: if $\alpha R \beta$ and $\alpha R \gamma$ then $\beta R \gamma$.
- Asymmetry: if $\alpha R \beta$ then not $\beta R \alpha$.
- Backward linearity: if $\alpha R \gamma$ and $\beta R \gamma$ then either $\alpha = \beta$ or $\alpha R \beta$ or $\beta R \alpha$.

Furthermore, properties linking two relations might also be imposed (e.g. one might want to require $R_2$ to be a subrelation of $R_1$: if $\alpha R_2 \beta$ then $\alpha R_1 \beta$).

Subsets of $\Omega$ are called *events*. Events represent propositions. The interpretation of events as propositions is obtained by means of the second component of modal logic: the *syntax* or formal language. The alphabet of the language consists of: (1) a finite or countable set $\mathbb{A}$ of atomic propositions (such as "the earth is flat"), (2) the connectives $\neg$ (for "not"), $\vee$ (for "or"), (3) the bracket symbols ( and ) and $m$ modal operators $\Box_1$, $\Box_2$, ..., $\Box_m$.[3] The connectives and the

[3] As is customary, we shall often omit the outermost brackets, e.g. we shall write $A \vee B$ instead of $(A \vee B)$, and use the following (metalinguistic) abbreviations: $A \wedge B$ for $\neg(\neg A \vee \neg B)$ (the symbol $\wedge$ stands for "and"), $A \rightarrow B$ for $(\neg A) \vee B$ (the symbol $\rightarrow$ stands for "if ... then ...") and $A \leftrightarrow B$ for $(A \rightarrow B) \wedge (B \rightarrow A)$ (the symbol $\leftrightarrow$ stands for "if and only if").

modal operators are used to form more complex sentences or *formulae.* The set $\mathbb{F}$ of formulae is obtained from the atomic propositions by closing with respect to negation, disjunction and the modal operators. That is, $\mathbb{F}$ is obtained recursively as follows: (i) for every atomic proposition $a \in \mathbb{A}$, $(a) \in \mathbb{F}$, (ii) if $A, B \in \mathbb{F}$ then $(\neg A) \in \mathbb{F}$, $(A \vee B) \in \mathbb{F}$ and, for every $j = 1, ..., m$, $(\Box_j A) \in \mathbb{F}$. The interpretation of the modal formula $\Box A$ depends on the context. Possible interpretations are: "the individual believes that $A$", "the individual knows that $A$", "it is always (i.e. at every possible future moment) going to be the case that $A$", etc.

The link between semantics and syntax is provided by the notion of *model.* Given a frame $\langle \Omega, R_1, ..., R_m \rangle$, a *model* $\mathcal{M}$ based on it is obtained by adding a *valuation* $V : \mathbb{A} \to 2^{\Omega}$ (where $2^{\Omega}$ denotes the set of subsets of $\Omega$) that associates with every atomic proposition $a$ the set of states at which $a$ is true. $\mathcal{M}, \omega \models A$ denotes that formula $A$ *is true at state* $\omega$ *in model* $\mathcal{M}$ and $\mathcal{M}, \omega \not\models A$ denotes that $A$ is false at $\omega$. For an atomic proposition $a$, $\mathcal{M}, \omega \models a$ if and only if $\omega \in V(a)$; furthermore, $\mathcal{M}, \omega \models \neg A$ if and only if $\mathcal{M}, \omega \not\models A$ and $\mathcal{M}, \omega \models (A \vee B)$ if and only if either $\mathcal{M}, \omega \models A$ or $\mathcal{M}, \omega \models B$. [4] Truth for the modal formula $\Box_j A$ is defined as follows:

$$\mathcal{M}, \alpha \models \Box_j A \quad \text{if and only if, for all } \beta \text{ such that } \alpha R_j \beta, \quad \mathcal{M}, \beta \models A.$$

Thus $\Box_j A$ is true at state $\alpha$ if and only if $A$ is true at every state which is $R_j$-accessible from $\alpha$.

A formula $A$ is *valid in model* $\mathcal{M}$ if it is true at every state, that is, if $\mathcal{M}, \omega \models A$ for all $\omega \in \Omega$.

**Remark 1.** *(a) If $A$ is valid in model $\mathcal{M}$ then, for every modal operator $\Box_j$, the formula $\Box_j A$ is valid in $\mathcal{M}$;*

*(b) If $A \to B$ is valid in model $\mathcal{M}$ then, for every modal operator $\Box_j$, the formula $\Box_j A \to \Box_j B$ is valid in $\mathcal{M}$.*

Often properties of the accessibility relation correspond to modal formulae. Examples of this correspondence are given in the following remark (cf. Chellas, 1984, p. 164; to simplify the notation we have dropped the subscript).

[4] It follows that $\mathcal{M}, \omega \models (A \wedge B)$ if and only if $\mathcal{M}, \omega \models A$ and $\mathcal{M}, \omega \models B$, and $\mathcal{M}, \omega \models (A \to B)$ if and only if $\mathcal{M}, \omega \models B$ whenever $\mathcal{M}, \omega \models A$.

**Remark 2.** *(a) Seriality of $R$ corresponds to $\Box A \to \neg\Box\neg A$ , that is, the following are equivalent: (1) $R$ is serial, (2) for every model $\mathcal{M}$ and for every formula $A$, the formula $\Box A \to \neg\Box\neg A$ is valid in $\mathcal{M}$.*
*(b) Reflexivity of $R$ corresponds to $\Box A \to A$.*
*(c) Transitivity of $R$ corresponds to $\Box A \to \Box\Box A$.*
*(d) Euclideanness of $R$ corresponds to $\neg\Box A \to \Box\neg\Box A$*

The modal logician is mainly concerned with proving soundness and completeness of a formal system with respect to the semantics. However, the focus of this paper is mainly on the validity of general claims about games and we shall ignore issues of completeness.

## 3. Epistemic logic and solution concepts

The view of game theory as a description of how *Homo rationalis* behaves in interactive situations has led to a large literature trying to identify for any game the strategies that might be chosen by rational and intelligent players who recognize each other's rationality and reasoning abilities. In this literature players' knowledge and beliefs play a central role. The "mutual recognition" of each other's rationality is modelled by means of the notion of *common belief.* Since this literature has been reviewed extensively (cf. Footnote 2), in this section we will just give a flavor of the type of results that have been proved.

The class of frames considered here are frames of the form $\langle \Omega, R_1, ..., R_n, R_* \rangle$ where $\{1, .., n\}$ is the set of players and the $(n+1)th$ relation $R_*$ is the transitive closure of $R_1 \cup ... \cup R_n$ [5].

The intended interpretation of the modal formula $\Box_i A$ $(i = 1, ..., n)$ is "player $i$ *believes* that $A$" while the interpretation of $\Box_* A$ (where $\Box_*$ is the modal operator associated with $R_*$) is "it is *common belief* that $A$". It is standard to require each $R_i$ $(i = 1, .., n)$ to be serial, transitive and euclidean, thus requiring individual beliefs to be consistent (if the individual believes $A$ then she does not believe $\neg A$: cf. Remark 2) and satisfy positive introspection (if the individual believes $A$ then she believes that she believes $A$) as well as negative introspection (if the individual does not believe $A$ then she believes that she does not believe $A$). Any

[5] That is, for all $\alpha, \beta \in \Omega$, $\alpha R_* \beta$ if and only if there is a sequence $\langle i_1, \cdots, i_m \rangle \in \{1, ..., n\}$ (the set of players) and a sequence $\langle \eta_0, \eta_1, \cdots, \eta_m \rangle$ in $\Omega$ (the set of states) such that: (i) $\eta_0 = \alpha$, (ii) $\eta_m = \beta$ and (iii) for every $k = 0, ..., m-1$, $\eta_k R_{i_{k+1}}(\eta_{k+1})$.

model based on such a frame validates also the following formulae, which capture the notion of common belief:[6]

$$\begin{aligned}&\Box_* A \to \Box_i A\\&\Box_* A \to \Box_i \Box_* A\\&\Box_*(A \to \Box_1 A \wedge ... \wedge \Box_n A) \to (\Box_1 A \wedge ... \wedge \Box_n A \to \Box_* A)\,.\end{aligned}$$

We shall restrict attention to finite non-cooperative games in strategic form[7], which are tuples $G = \langle N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N}\rangle$, where $N = \{1, 2, \ldots, n\}$ is the set of players, $S_i$ is the finite set of strategies for player $i$ and $u_i : S \to \mathbb{R}$ (where $S = S_1 \times \ldots \times S_n$ and $\mathbb{R}$ is the set of real numbers) is player $i$'s von Neumann Morgenstern payoff (or utility) function. This definition provides only a partial description of the interactive situation, in that it determines the choices that are available to the players and their preferences, but does not specify the players' beliefs about each other or their actual choices. The notion of model of a game provides a way of completing the description.

Since the objective of this literature is to consider the implications of common belief of rationality, the formal language we consider needs to include atomic propositions such as "player $i$ is rational". Furthermore, we need atomic propositions that refer to properties of strategy profiles. Given that the purpose of this section is merely to illustrate the type of results obtained in this literature, we will limit ourselves to the property of survival of iterative deletion of strictly dominated strategies. Recall that a probability distribution over $S_i$ is called a *mixed* strategy of player $i$, whereas the elements of $S_i$ are called player $i$'s *pure* strategies. If $\nu_i \in \Delta(S_i)$ (where $\Delta(S_i)$ denotes the set of probability distributions over $S_i$) and $s_i \in S_i$, we denote by $\nu_i(s_i)$ the probability assigned to $s_i$ by $\nu_i$. Let $S_{-i} = S_1 \times ... \times S_{i-1} \times S_{i+1} \times ... \times S_n$ denote the set of pure-strategy profiles of the players other than $i$. A pure strategy $s_i \in S_i$ of player $i$ is *strictly dominated by* $\nu_i \in \Delta(S_i)$ if, for all $s_{-i} \in S_{-i}$, $\sum\limits_{x \in S_i} \nu_i(x)\; u_i(x, s_{-i}) > u_i(s_i, s_{-i})$. Given a game $G$, let $G^0 = G$ and, for $k \geq 1$, let $G^k$ be the game obtained by deleting

[6] From $\Box_* A \to \Box_j A$ one obtains (cf. Remark 1) $\Box_i \Box_* A \to \Box_i \Box_j A$. From this and $\Box_* A \to \Box_i \Box_* A$ one obtains (by standard propositional logic) the validity of $\Box_* A \to \Box_i \Box_j A$. This argument can be repeated any number of times to yield the conclusion that if $A$ is common belief then everybody believes that everybody believes ... (any number of times) that everybody believes that $A$. The converse can also be proved. See, for example, Bonanno (1996), Lismont (1993) and Lismont and Mongin (1994).

[7] The epistemic analysis of solution concepts for *extensive* games has also been carried out: see, for example, Battigalli and Siniscalchi (1999).

all the pure strategies of all the players that are strictly dominated in $G^{k-1}$. Let $G^{\infty} = \bigcap_{k=1}^{\infty} G^k$ be the game obtained by applying this iterative elimination procedure (by finiteness of $S$, there is some $m$ such that $G^{\infty} = G^m$). Let $S^{\infty}$ be set of strategy profiles in $G^{\infty}$ (that is, the set of strategy profiles of $G$ that survive the iterative deletion of strictly dominated strategies).

By a *game language* we mean a language that contains the following atomic propositions:

| Symbol | Intended interpretation |
|---|---|
| $r_i$ | player $i$ is rational |
| $s^{\infty}$ | the strategy profile played belongs to the set $S^{\infty}$. |

Let $r$ denote the formula $r_1 \wedge ... \wedge r_n$ whose interpretation is "all the players are rational".

A *probabilistic frame* is a tuple $\langle \Omega, R_1, ..., R_n, R_*, P_1, ..., P_n \rangle$ where:

- $\Omega$ is a finite set of states,
- each relation $R_i$ $(i = 1, ..., n)$ is serial, transitive and euclidean, and $R_*$ is the transitive closure of $R_1 \cup ... \cup R_n$, and
- $P_i$ is a full-support probability distribution on $\Omega$ (that is, $P_i(\omega) > 0$, $\forall \omega \in \Omega$).

The introduction of a full-support probability distribution $P_i$ is just a notationally convenient way of assigning probabilistic beliefs to each player. $P_i$ in itself is of no significance; what matters is player $i$'s belief at each state $\alpha$, denoted by $p_{i,\alpha}$, which is obtained by conditioning $P_i$ on $R_i(\alpha) = \{\omega \in \Omega : \alpha R_i \omega\}$:

$$p_{i,\alpha}(\omega) = \begin{cases} \frac{P_i(\omega)}{\sum\limits_{\omega' \in R_i(\alpha)} P_i(\omega')} & if \;\; \omega \in R_i(\alpha) \\ 0 & if \;\; \omega \notin R_i(\alpha). \end{cases}$$

As explained in the previous section, in general a model is obtained by adding to a frame a valuation $V$ that associates with every atomic proposition the set of states at which the proposition is true. In order to obtain a model of a particular game $G$, besides the valuation $V$ we also need to add a function $\sigma = (\sigma_1, ..., \sigma_n) : \Omega \rightarrow S$ that associates with every state the *pure* strategy profile played at that state. We call the pair $(V, \sigma)$ a *G-valuation* and the model so obtained a *G-model* if the valuation $(V, \sigma)$ satisfies the following restrictions (which realize the intended interpretation of $r_i$ and $s^{\infty}$; let $\sigma_{-i} = (\sigma_1, ..., \sigma_{i-1}, \sigma_{i+1}, ..., \sigma_n)$):

**1.** $\alpha \in V(r_i)$ if and only if: (a) player $i$ has no uncertainty as to the strategy he himself is playing, that is,

$$if \quad \alpha R_i \beta \quad then \quad \sigma_i(\beta) = \sigma_i(\alpha)$$

and (b) $\sigma_i(\alpha)$ (player $i$'s strategy at $\alpha$) maximizes $i$'s expected utility given his beliefs, that is,

$$\sum_{\omega \in R_i(\alpha)} u_i\left(\sigma_i(\alpha), \sigma_{-i}(\omega)\right)\ p_{i,\alpha}(\omega) \geq \sum_{\omega \in R_i(\alpha)} u_i\left(x, \sigma_{-i}(\omega)\right)\ p_{i,\alpha}(\omega)\,, \quad \forall x \in S_i.$$

**2.** $\alpha \in V(s^\infty)$ if and only if $\sigma(\alpha) \in S^\infty$, that is, if and only if the strategy profile played at $\alpha$ survives the iterative deletion of strictly dominated strategies.

As usual, the formula $\Box_i A$ is true at state $\alpha$ ($\mathcal{M}, \alpha \models \Box_i A$) if and only if $A$ is true at every state $\beta$ such that $\alpha R_i \beta$. Thus, in this context, the interpretation of $\Box_i A$ is "player $i$ believes (with probability 1) that $A$". Similarly, $\mathcal{M}, \alpha \models \Box_* A$ if and only if $\mathcal{M}, \beta \models A$ for every $\beta$ such that $\alpha R_* \beta$ and the interpretation of $\Box_* A$ is "it is common belief that $A$".

The following result provides an epistemic characterization of the procedure of iterative deletion of strictly dominated strategies. The essence of this result is due to Bernheim (1984) and Pearce (1984), although they did not make use of the apparatus of epistemic logic. The first epistemic characterization was provided by Tan and Werlang (1988) using a universal type space (rather than modal logic). The formulation which is closest to the one given below is that of Stalnaker (1994), but it was implicit in Brandenburger and Dekel (1987).

**Proposition 3.1.** *Let $G$ be a finite strategic form game. Then the following formula is valid in every model of $G$:*

$$\Box_* r \to s^\infty.$$

That is, if there is common belief that all the players are rational, then the strategy profile actually played is one that survives the iterative deletion of strictly dominated strategies.

The literature on the epistemic foundations of game theory has dealt with several other solution concepts: Nash equilibrium, correlated equilibrium, backward and forward induction, etc. Since this literature has been reviewed elsewhere (see Footnote 2), we shall turn to the alternative view of solution concepts, namely solutions as recommendations to the players.

## 4. Solutions as consistent recommendations

Although the main focus of this section will be extensive (or dynamic) games, to maintain continuity with the previous section we shall start with finite strategic-form games and provide a (straightforward) interpretation of Nash equilibrium as a recommendation. As before, a frame is a tuple $\langle \Omega, R_1, ..., R_n, R_* \rangle$, where $\Omega$ is a set of states and the index $i = 1, ..., n$ refers to the players. However, the interpretation that we want to establish for $\alpha R_i \beta$ is no longer "for player $i$ state $\beta$ is epistemically accessible from (or an epistemic alternative to) state $\alpha$" but rather "from state $\alpha$ player $i$ can *unilaterally* bring about state $\beta$". Thus $R_i$ does not capture the reasoning or epistemic state of player $i$ but rather the notion of what player $i$ is *able to do*. Let $R^T$ be the transitive closure of $\bigcup_{i=1}^{n} R_i$. Thus if $\alpha R^T \beta$ then from state $\alpha$ the players can collectively bring about state $\beta$ by a series of individual actions. The relation $R_*$ is now assumed to be a *subrelation* of $R^T$. It is no longer meant to capture the epistemic notion of common belief; instead it will be interpreted as expressing the theory's *recommendation* to the players. Thus the intended interpretation of $\alpha R_* \beta$ is "at state $\alpha$ it is recommended that state $\beta$ be reached".

As before, the formal language needs to contain atomic propositions that can be interpreted as statements about the game. We include the following atomic propositions ($p_i, q_i \in \mathbb{Q}$, where $\mathbb{Q}$ is the set of rational numbers):

| symbol | intended interpretation |
|---|---|
| $(u_i = p_i)$ | "player $i$'s utility (or payoff) is $p_i$" |
| $(q \leq p)$ | "the rational number $q$ is less than or equal to the rational number $p$" |
| $Nash$ | "the pure strategy profile played is a Nash equilibrium"[8]. |

As in the previous section, given a strategic-form game $G$, a $G$-valuation is a pair $(V, \sigma)$ where the function $V$ associates with every atomic proposition $a$ the set of states where $a$ is true and $\sigma$ is a function that associates with every state

[8] A strategy profile $s \in S$ is a *Nash equilibrium* if $\forall i = 1, ..., n$, $\forall x \in S_i$, $u_i(s) \geq u_i(x, s_{-i})$.

a profile of pure strategies. The valuation is now assumed to satisfy the following requirements:

1. $\alpha R_i \beta$ if and only if $\sigma_{-i}(\beta) = \sigma_{-i}(\alpha)$ (this requirement captures the notion of "bringing about *unilaterally*");

2. if $a$ is an atomic proposition of the form $(q \leq p)$ with $p, q \in \mathbb{Q}$ then $V(a) = \Omega$ if $q \leq p$ and $V(a) = \emptyset$ otherwise;

3. $\alpha \in V(u_i = p_i)$ if and only if $u_i(\sigma(\alpha)) = p_i$;

4. $\alpha \in V(Nash)$ if and only if $\sigma(\alpha)$ is a Nash equilibrium of $G$.

In this context, the interpretation of $\Box_i A$ is "no matter what unilateral action player $i$ takes, $A$ is true" whereas $\Box_* A$ can be interpreted as "it is recommended that $A$".

To simplify the notation, we shall write ${}_\wedge(u_i = p_i)$ for $(u_1 = p_1) \wedge ... \wedge (u_n = p_n)$ and ${}_\wedge\Box_i A$ for $\Box_1 A \wedge ... \wedge \Box_n A$.

The following lemma is a straightforward translation of the definition of Nash equilibrium in the formal language.

**Lemma 4.1.** *Let $G$ be a finite strategic-form game. Then the following formula is valid in every model of $G$:*
$${}_\wedge(u_i = p_i) \;\wedge\; {}_\wedge\Box_i((u_i = q_i) \to (q_i \leq p_i)) \quad \leftrightarrow \quad {}_\wedge(u_i = p_i) \;\wedge\; Nash.$$

That is, if player $i$'s payoff is $p_i$ and, no matter what unilateral action player $i$ takes, it is the case that if his payoff is $q_i$ then $q_i$ is not greater than $p_i$, then the strategy profile actually played is a Nash equilibrium and *vice versa.*

Using the above lemma it is easy to prove the following.

**Proposition 4.2.** *Let $G$ be a finite strategic-form game. Then the following formula is valid in every model of $G$:*
$$\Box_* \left({}_\wedge(u_i = p_i) \to {}_\wedge\Box_i((u_i = q_i) \to (q_i \leq p_i)) \right. \quad \to \quad \Box_*(Nash).$$

That is, if it is recommended that the game be played in such a way that whenever player $i$'s payoff is $p_i$ then, no matter what unilateral action player $i$ takes, it is the case that if his payoff is $q_i$ then $q_i$ is not greater than $p_i$, then the recommendation is that a Nash equilibrium be played.

We now turn to the more interesting case of extensive-form games. We shall restrict attention to games with perfect information and our purpose is to provide a characterization of backward induction in terms of the notion of internal consistency of a recommendation.

Recall that a *rooted tree* is a pair $\langle \Omega, \rightarrowtail \rangle$ where $\Omega$ is a set of *nodes* and $\rightarrowtail$ is a binary relation on $\Omega$ (if $\omega \rightarrowtail \omega'$ we say that $\omega$ *immediately precedes* $\omega'$ or that $\omega'$ *immediately succeeds* $\omega$) satisfying the following properties:

1. there is a unique node $\omega_0$ with no immediate predecessors; it is called the *root*;

2. for every node $\omega \in \Omega \backslash \{\omega_0\}$ there is a unique path from $\omega_0$ to $\omega$, that is, there is a unique sequence $\langle x_1, ..., x_m \rangle$ in $\Omega$ with $x_1 = \omega_0$, $x_m = \omega$, and, for every $j = 1, ...m - 1$, $x_j \rightarrowtail x_{j+1}$.

Given a rooted tree $\langle \Omega, \rightarrowtail \rangle$, a *terminal node* is an $\omega \in \Omega$ which has no immediate successors. Let $Z \subseteq \Omega$ denote the set of terminal nodes. If $\Omega$ is finite then $Z \neq \emptyset$.

**Definition 4.3.** *A finite extensive form with perfect information is a tuple $\langle \Omega, \rightarrowtail, N, \iota \rangle$ where $\langle \Omega, \rightarrowtail \rangle$ is a finite rooted tree, $N = \{1, ..., n\}$ is the set of* players *and $\iota : \Omega \backslash Z \to N$ is a function that associates with every non-terminal or* decision *node the player who* moves *at that node. If $i = \iota(\omega)$ and $\omega \rightarrowtail \omega'$ we say that the pair $(\omega, \omega')$ is a* choice *of player $i$ at node $\omega$. Given a finite extensive form with perfect information one obtains a perfect information* game *by adding, for every $i \in N$, a* payoff *or* utility *function $u_i : Z \to \mathbb{Q}$ (where $Z$ is the set of terminal nodes and $\mathbb{Q}$ is the set of rational numbers).*

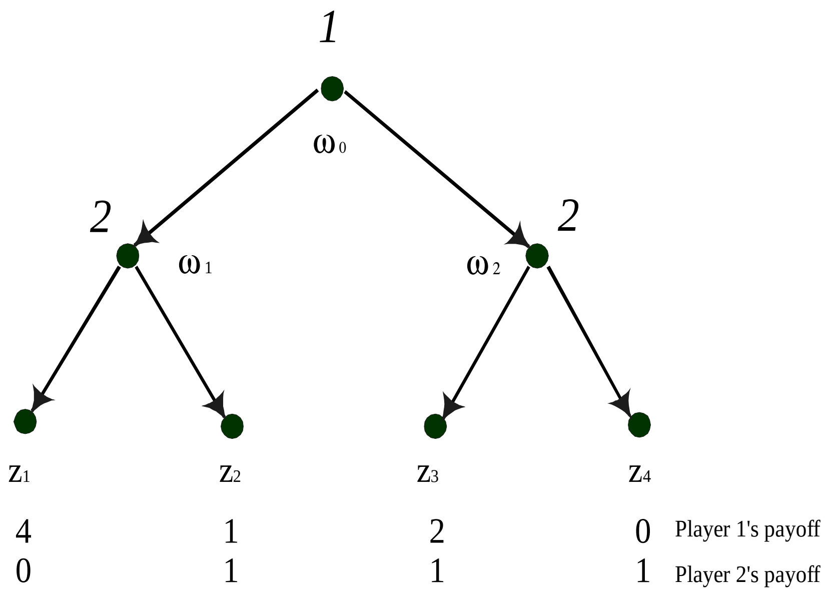

*Figure 1*

Figure 1 shows a perfect information game with two players. The vector $(x_1, x_2)$ written next to a terminal node $z$ represents the payoff vector $(u_1(z), u_2(z))$ and there is an arrow from $\omega$ to $\omega'$ if and only if $\omega \rightarrowtail \omega'$. For every decision node $\omega$, the corresponding player $\iota(\omega)$ is written next to it.

A well-known procedure for solving a perfect information game is the *backward induction* algorithm first used by Zermelo (1913) for the game of chess. The algorithm starts at the end of the game and proceeds backwards towards the root:

1. Start from a decision node $\omega$ whose immediate successors are only terminal nodes (e.g. node $\omega_1$ in Figure 1) and select one choice that maximizes the utility of player $\iota(\omega)$ (in the example of Figure 1, at $\omega_1$ player 2 should make the choice that leads to node $z_2$ since it gives her a payoff of 1 rather than 0, which is the payoff that she would get if the play proceeded to node $z_1$). Turn $\omega$ into a terminal node by deleting the immediate successors of $\omega$ and assigning to $\omega$ the payoff vector associated with the selected choice.

2. Repeat until all the decision nodes have been exhausted.

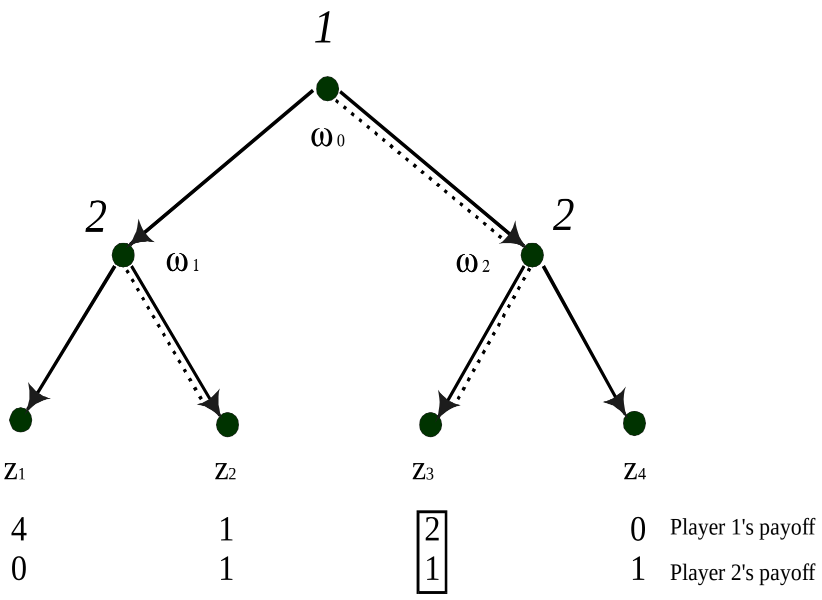

*Figure 2*

Figure 2 shows a possible outcome of the backward induction algorithm for the game of Figure 1. The choices selected by the algorithm are shown as dotted lines next to the corresponding arrows.

The backward induction algorithm may yield more than one solution. Multiplicity may arise if there are players who have more than one utility-maximizing choice. For example, in the game of Figure 1 at $\omega_2$ both choices are optimal for Player 2. The selection of choice $(\omega_2, z_3)$ leads to the solution shown in Figure 2, while the selection of choice $(\omega_2, z_4)$ leads to a different solution shown in Figure 3.

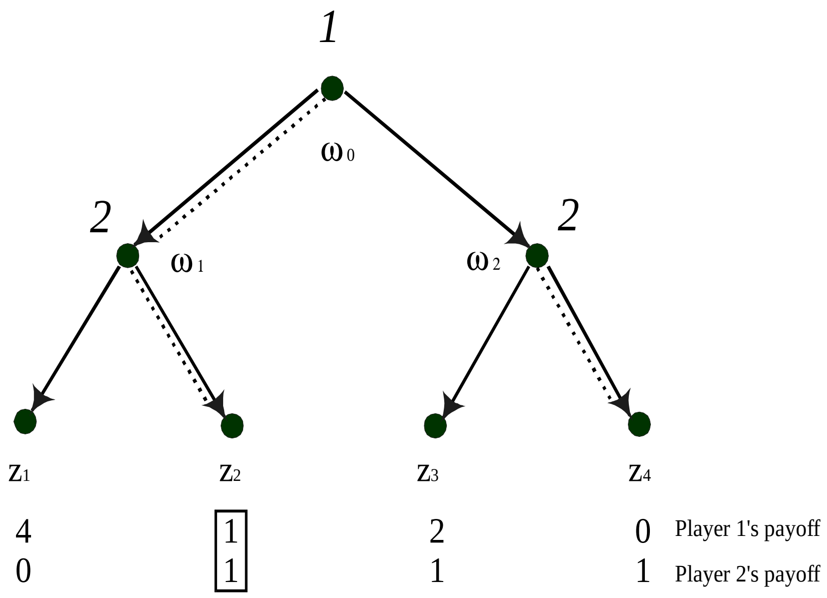

*Figure 3*

**Definition 4.4.** *A perfect information game is generic if no player is indifferent between any two terminal nodes, that is,* $\forall i \in N$, $\forall z, z' \in Z$ *if* $u_i(z) = u_i(z')$ *then* $z = z'$.

**Remark 3.** *In a generic game the backward induction algorithm yields a unique solution.*

For simplicity, in the following discussion we shall restrict attention to generic games.

The relationship between an extensive *form* with perfect information and a perfect information *game* is similar to the relationship between a frame and a model. Indeed an extensive form can be viewed as a frame, as follows. First of all, recall that a frame is a tuple $\langle \Omega, R_1, ..., R_n, R_* \rangle$, where $\Omega$ is a set of states and the remaining objects are binary relations on $\Omega$. In this section the intended interpretation of $\alpha R_i \beta$ is "at node $\alpha$ player $i$ can bring about node $\beta$", while the relation $R_*$ is intended to represent the theory's *recommendation* to the players:

$\alpha R_* \beta$ is interpreted as "at node $\alpha$ it is recommended that the sequence of actions be taken that leads from $\alpha$ to $\beta$".

**Definition 4.5.** *Given a finite extensive form with perfect information $\langle \Omega, \rightarrowtail, N, \iota \rangle$, we say that $\langle \Omega, R_1, ..., R_n, R_* \rangle$ is a compatible frame if:*

1. For all $i \in N$ and $\omega, \omega' \in \Omega$, $\omega R_i \omega'$ if and only if $i = \iota(\omega)$ and $\omega \rightarrowtail \omega'$;
2. $R_*$ is a subrelation of $R^T$ (the transitive closure of $\bigcup_{i=1}^{n} R_i$)[9] satisfying the following properties:
   (a) $R_*$ is transitive,
   (b) if $\omega$ is a decision node then $\omega R_* \omega'$ for some $\omega'$,
   (c) if $\omega_1 R_* \omega_3$, $\omega_1 R^T \omega_2$ and $\omega_2 R^T \omega_3$ then $\omega_1 R_* \omega_2$ and $\omega_2 R_* \omega_3$.

Property (a) requires the recommendation to be "forward-consistent", in the sense that if the recommendation is to go from $\omega_1$ to $\omega_2$ and, once $\omega_2$ is reached, the recommendation is then to go to $\omega_3$, then going to $\omega_3$ must be a recommendation at the initial node $\omega_1$. Property (b) requires that, as long as a node has a successor, then some recommendation be made. Note, however, that this is not a stringent requirement, since it does not rule out the recommendation "do anything", that is, it does not rule out the case where $\omega R_* \omega'$ if and only if $\omega R^T \omega'$. Finally, property (c) is a requirement of coherence of the recommendation with the tree structure of the game: if a recommended path of the game leads from node $\omega_1$ to node $\omega_3$ ($\omega_1 R_* \omega_3$) and $\omega_2$ is a node that belongs to that path ($\omega_1 R^T \omega_2$ and $\omega_2 R^T \omega_3$) then a recommendation at $\omega_1$ must be that $\omega_2$ be reached and a recommendation at $\omega_2$ must be that $\omega_3$ be reached. This is a natural requirement, since in a tree the path from $\omega_1$ to $\omega_3$ is unique and therefore one *must* go through node $\omega_2$ when following the recommended path from $\omega_1$ to $\omega_3$.

For example, for the extensive form of Figure 1, the following is a compatible frame: $R_1 = \{(\omega_0, \omega_1), (\omega_0, \omega_2)\}$, $R_2 = \{(\omega_1, z_1), (\omega_1, z_2), (\omega_2, z_3), (\omega_2, z_4)\}$, $R_* = \{(\omega_0, \omega_1), (\omega_1, z_1), (\omega_1, z_2), (\omega_0, z_1), (\omega_0, z_2), (\omega_2, z_3)\}$. This example illustrates the fact that Definition 4.5 does *not* require the recommendation to select a unique path to a terminal node out of every decision node. It is quite possible for $R_*$ to select multiple paths, in which case the recommendation would be "do either this or that". The extreme case is where $R_* = R^T$, that is, every path out of every

[9] Thus $\alpha R^T \beta$ if and only if there is a path in the tree from node $\alpha$ to node $\beta$. Note that, besides transitivity, the relation $R^T$satisfies asymmetry and backward linearity (cf. Section 2).

decision node is recommended, so that the recommendation is "do anything". Another example of a frame which is compatible with the extensive form of Figure 1 is $R_1$ and $R_2$ as above and $R_* = \{(\omega_0, \omega_1), (\omega_1, z_2), (\omega_0, z_2), (\omega_2, z_4)\}$, which is the transitive closure of the the backward-induction relation shown in Figure 3 as dotted lines.

**Definition 4.6.** *Given a generic extensive game with perfect information, the backward-induction algorithm determines for every decision node a unique immediate successor, thus giving rise to a relation on the set of nodes $\Omega$. Call it the backward-induction relation. We say that the relation $R_*$ is the backward-induction recommendation if it is the transitive closure of the backward-induction relation. It is easy to check that the backward-induction recommendation satisfies properties (a)-(c) of Definition 4.5.*

As explained before, to view a perfect information *game* as a model all we need to do is include in the set of atomic propositions sentences of the form $(u_i = p_i)$ whose intended interpretation is "player $i$'s utility (or payoff) is $p_i$"; furthermore we need to add the standard ordering of the rational numbers by means of sentences of the form $(q \leq p)$ whose intended interpretation is "the rational number $q$ is less than or equal to the rational number $p$". A *game language* is a language obtained as explained in Section 2 from a set $\mathbb{A}$ of atomic propositions that includes sentences of the form $(u_i = p_i)$ and $(q \leq p)$.

**Definition 4.7.** *Let $G$ be a perfect information game and $\mathcal{F}$ be a compatible frame (cf. Definition 4.5). A game model is a model based on $\mathcal{F}$ (cf. Section 2) obtained in a game language by adding to $\mathcal{F}$ a valuation $V : \mathbb{A} \to 2^{\Omega}$ satisfying the following properties:*

- if $a \in \mathbb{A}$ is of the form $(q \leq p)$ with $p, q \in \mathbb{Q}$ then $V(a) = \Omega$ if $q \leq p$ and $V(a) = \emptyset$ otherwise
- if $a \in \mathbb{A}$ is of the form $(u_i = p_i)$ then $V(a) = \{z \in Z : u_i(z) = p_i\}$.

Thus if $\mathcal{M}$ is a game model then, $\forall \omega \in \Omega$, $\mathcal{M}, \omega \models (q \leq p)$ if $q$ is less than or equal to $p$ and $\mathcal{M}, \omega \models \neg(q \leq p)$ otherwise; furthermore, $\mathcal{M}, \omega \models (u_i = p_i)$ if $\omega$ is a terminal node with $u_i(\omega) = p_i$ and $\mathcal{M}, \omega \models \neg(u_i = p_i)$ if $\omega$ is either a decision node or a terminal node with $u_i(\omega) \neq p_i$. The valuation of the other formulae is as explained in Section 2.

In this context, the interpretation of the modal formula $\Box_i A$ is "no matter what action player $i$ takes, it will be the case that $A$" whereas $\Box_* A$ can be interpreted as "if the recommendation is followed then it will be the case that $A$". Hence $\neg\Box_*\neg A$ is interpreted as "it is possible, according to the recommendation, that $A$".

Consider the following axiom scheme:

$$\neg\Box_*\neg(u_i = p_i) \;\rightarrow \Box_i \left(((u_i = q_i) \vee \neg\Box_*\neg(u_i = q_i)) \;\rightarrow\; q_i \leq p_i\right). \qquad \text{(IC)}$$

(IC) says that if, according to the recommendation, it is possible that player $i$'s payoff is $p_i$, then, no matter what action player $i$ takes, it will be the case that if player $i$'s payoff is, or is recommended to be, $q_i$ then $q_i$ is not greater than $p_i$. In other words, (IC) says that if the recommendation is that the game be played in such a way that player $i$ gets a payoff of $p_i$ then it is not possible for player $i$ to take an action after which either his payoff is greater than $p_i$ or the recommendation is that the game be played in such a way that player $i$ gets a payoff greater than $p_i$. Thus (IC) can be viewed as expressing a notion of *internal consistency* (hence the name IC) of a recommendation, in the sense that no player can increase his payoff by deviating from the recommendation, *using the recommendation itself to predict his future payoff after the deviation.*

The following proposition is an adaptation of a result proved in Bonanno (2001b).[10]

**Proposition 4.8.** *Let $G$ be a generic perfect information game and $\mathcal{F} = \langle \Omega, R_1, ..., R_n, R_* \rangle$ a compatible frame (cf. Definition 4.5). Then the following are equivalent:*
*(1) $R_*$ is the backward induction recommendation,*
*(2) axiom (IC) is valid in every model of $G$ based on $\mathcal{F}$ (cf. Definition 4.7).*

While the recent debate on backward induction has focused entirely on whether backward induction can be validly derived from the hypothesis of common belief in rationality, the above proposition shows that there are also characterizations of it that are independent both of the notion of (Bayesian) rationality and of epistemic hypotheses about the players.

[10] Bonanno (2001b) applies to extensive games the notion of prediction developed in Bonanno (2001a) within the context of temporal modal logic.

## 5. Conclusion

The main purpose of this paper was not to introduce new results but to show that the same formal tool, namely the apparatus of modal logic, can be used to model two conceptually very different views of game theory. One is the view that game-theoretic solution concepts capture the reasoning of rational players, that is, how *Homo rationalis* reaches a decision on how to play the game. The other view is that solution concepts are not descriptive but normative: they provide advice to the players on how to play the game. The advice acknowledges the goals of the players, as represented by their payoffs, and it does so in a way which is not self-defeating. While the first view has been investigated extensively in the literature, little attention has been devoted to the second view of game theory.

Since this paper was entirely concerned with the use of modal logic in the analysis of games, it seems appropriate to conclude with a discussion of the usefulness of this approach. The most important advantages of using modal logic are the following.

1. The tools of modal logic have enriched the game-theoretic language by making it possible to express concepts that were previously either informally or vaguely claimed to be captured by a solution concept. A good example of this is the notion of common belief in rationality and its relationship to the procedure of iterative deletion of strictly dominated strategies (Proposition 3.1). Another example is given by the relationship between the notion of internal consistency and the backward-induction algorithm (Proposition 4.8).

2. Once concepts, such as common belief, are modeled explicitly, new questions arise in a natural way. For example, the fundamental difference between knowledge and belief is that, while knowledge is veridical (only true facts can be known), beliefs can be mistaken (it is possible to believe something which is false). Thus one can ask whether by ruling out, at some level, incorrect beliefs one can further restrict the strategy profiles that are compatible with common belief in rationality. This question led Stalnaker (1994) to uncover a new solution concept, "strong rationalizability". The strongly rationalizable strategies are a proper subset of the ones that survive the iterative deletion of strictly dominated strategies and are those that are consistent with common belief in rationality when (i) it is common belief that no player has incorrect

beliefs and (ii) collectively players are in fact correct in their beliefs (see Bonanno and Nehring, 1998).

3. As Bacharach (1994, p. 21) notes,

   > "Game theory is full of deep puzzles, and there is often disagreement about proposed solutions to them. The puzzlement and disagreement are neither empirical nor mathematical but, rather, concern the meanings of fundamental concepts ('solution', 'rational','complete information') and the soundness of certain arguments (that solutions must be Nash equilibria, that rational players defect in Prisoner's Dilemmas, that players should consider what would happen in eventualities which they regard as impossible). Logic appears to be an appropriate tool for game theory both because these conceptual obscurities involve notions such as reasoning, knowledge and counterfactuality which are part of the stock-in-trade of logic, and because it is a prime function of logic to establish the validity or invalidity of disputed arguments".

A good example of a disputed argument in game theory is whether backward induction in perfect-information games can be derived from the hypothesis of common belief in (or knowledge of) rationality. There are those (e.g. Aumann, 1995) who claim that the answer is positive and those (e.g. Stalnaker, 1998) who claim the opposite. In a recent contribution Halpern (2001) attempts to clarify the debate by highlighting a difference in the interpretation of the counterfactuals involved in evaluating players' rationality at unreached nodes in the game tree.

The main purpose of this paper was to point out that the same clarifying role that modal logic has played in the rationality-based interpretation of solution concepts can also be played in the alternative interpretation based on the notion of recommendation. The epistemic logic approach is in a sense "internal" to the players, in that it tries to capture explicitly their reasoning and their mutual recognition of each other's Bayesian rationality. The approach discussed in Section 4, on the other hand, is "external" to the players: it deals entirely with the notion of what it means for a recommendation to acknowledge the goals of the players (as expressed by their payoffs) in a consistent way. The notion of internal consistency used in Proposition 4.8 to characterize backward induction captures in an explicit way the informal notion that a recommendation "should not be a self-destroying prophecy which creates an incentive to deviate for those who believe in it" (Selten, 1985, p. 79). Given the conceptually very different nature of the two approaches, one could not easily deduce from the usefulness of modal logic in the epistemic approach that modal logic would prove to be the appropriate tool of analysis also in the normative interpretation of game theory.

## References

[1] Aumann, Robert (1985). What is game theory trying to accomplish?, in: K. Arrow and S. Honkapohja (Eds.), *Frontiers of economics*, Basil Blackwell, Oxford, 28-76.

[2] Aumann, Robert (1995). Backward induction and common knowledge of rationality, *Games and Economic Behavior,* **8**, 6-19.

[3] Bacharach, Michael (1994). The epistemic structure of a theory of a game, *Theory and Decision,***37**, 7-48.

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

[5] Battigalli, Pierpaolo and Marciano Siniscalchi (1999). Hierarchies of conditional beliefs and interactive epistemology in dynamic games, *Journal of Economic Theory*, **88**, 188-230.

[6] Bernheim, Douglas (1984). Rationalizable strategic behavior, *Econometrica,* **52**, 1002-1028.

[7] Bonanno, Giacomo (1996). On the logic of common belief, *Mathematical Logic Quarterly*, **42**, 305-311.

[8] Bonanno, Giacomo (2001a). Prediction in branching time logic, *Mathematical Logic Quarterly*, **47**, 239-247.

[9] Bonanno, Giacomo (2001b). Branching time, perfect information games and backward induction, *Games and Economic Behavior*, **36**, 57-73.

[10] Bonanno, Giacomo and Klaus Nehring (1998). On Stalnaker's notion of strong rationalizability and Nash equilibrium in perfect information games, *Theory and Decision*, **45**, 291-295.

[11] Brandenburger, Adam and Eddie Dekel (1987). Rationalizability and correlated equilibria, *Econometrica,* **55,** 1391-1402.

[12] Chellas, Brian (1984). *Modal logic: an introduction,* Cambridge University Press.

[13] Dekel, Eddie and Faruk Gul (1997). Rationality and knowledge in game theory, in: D. Kreps and K. Wallis (Eds.), *Advances in Economics and Econometrics*, Cambridge University Press, 87-172.

[14] Greenberg, Joseph (1990), *The theory of social situations*, Cambridge University Press.

[15] Halpern, Joseph (2001). Substantive rationality and backward induction, *Games and Economic Behavior*, **37**, 425-435.

[16] Lismont, Luc (1993). La connaissance commune en logique modale, *Mathematical Logic Quarterly*, **39**, 115-130.

[17] Lismont, Luc and Philippe Mongin (1994). On the logic of common belief and common knowledge, *Theory and Decision*, **37**, 75-106.

[18] Nash, John (1951). Non-cooperative games, *Annals of Mathematics*, 44, 286-295.

[19] Pearce, David (1984). Rationalizable strategic behavior and the problem of perfection, *Econometrica,* **52**, 1029-1050.

[20] Selten, Reinhardt (1985). Comment, in: K. Arrow and S. Honkapohja (Eds.), *Frontiers of economics*, Basil Blackwell, 77-87.

[21] Stalnaker, Robert (1994). On the evaluation of solution concepts, *Theory and Decision*, **37**, 49-74.

[22] Stalnaker, Robert (1998). Belief revision in games: forward and backward induction, *Mathematical Social Sciences,* **36**, 31-56.

[23] Tan, T. and S. Werlang (1988). The Bayesian foundation of solution concepts of games, *Journal of Economic Theory,* **45**, 370-391.

[24] von Neumann, John and Oscar Morgenstern (1947). *Theory of games and economic behavior*, Princeton University Press.

[25] Zermelo, E. (1913), Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels, in: E.W. Hobson and A.E.H. Love, Eds., *Preceedings of the Fifth International Congress of Mathematicians*, Vol. II, Cambridge University Press, 501-504.