Behavioural versus Habitual Rationality and Backward Induction *

Thorsten Clausing

Leipzig Graduate School of Management, Jahnallee 59, D-04109 Leipzig, Germany
clausing@microec.hhl.de

Abstract. The problem of finding sufficient conditions for backward induction in games of perfect information is analysed in a syntactic framework with subjunctive conditionals. The structure of the game is described by a logical formula. Two different rationality conditions are formulated, which are called behavioural and habitual rationality. While common knowledge of the former and the structure of the game does not imply backward induction, higher level knowledge of the latter and the structure of the game does. It is shown that similar results can be proven with non-introspective belief instead of knowledge.

1 Introduction

Does common knowledge of rationality imply backward induction in generic games of perfect information, i.e. in games of perfect information where no player gets the same payoff at two different terminal nodes? Even though by now a substantial number of papers dealing with this seemingly simple question has been published (for an overview, cf. Dekel and Gul, 1996), there is still widespread disagreement as to its correct answer. In this literature, it is usually taken as understood that there is also common knowledge of the structure of the game. The approach advocated in this paper tries to shed more light on the backward induction question by explicitly taking care of knowledge of the game and distinguishing between different concepts of rationality. To this end, a straightforward combination of standard systems of epistemic and conditional logic is used. This syntactic framework allows to express subjunctive conditionals. A subjunctive conditional is a statement of the form "if ϕ were true, then ψ would be true", which is to be interpreted as saying that in a hypothetical world where ϕ is true, but that is otherwise as similar to the actual world as possible, ψ would be true. Unlike the more familiar material implication $\phi \Rightarrow \psi$, a subjunctive conditional $\phi \hookrightarrow \psi$ may thus be false even though its antecedent ϕ is false. Note that many important game theoretic notions are based on conditional statements of this kind. Thus in describing a player's strategy in an extensive form game, one says which moves he would make if certain nodes were reached, and the fact that

* I would like to thank an anonymous referee and participants of the International Conference on Logic, Game Theory and Social Choice in Oisterwijk (The Netherlands) for helpful comments.

some node v will not be reached does not mean that the statement "if v were reached, move X would be played" is true for all possible moves X at v. In the same way, a description of the structure of the game tells you which payoffs the players would get if a certain play of the game obtained, independently of whether it will indeed obtain. Therefore the use of conditional logic appears as a rather natural choice of formalism.

The remainder of the paper is organised as follows. In the next section, I will give a detailed presentation of the logic that will be employed. Section 3 shows how the structure of a given game of perfect information can be described in this logic. Section 4 introduces two different notions of rationality and contains the main results on backward induction. In section 5, knowledge is replaced by belief, and the final section compares the approach of this paper to the relevant literature.

2 Conditional Epistemic Logic

To define the syntax of conditional epistemic logic, call X_v a move formula and $[\pi_i = x]$ a payoff formula with the intended interpretations "move X will be played at node v" and "player i will receive a payoff of x". For a given generic game of perfect information Γ, let $\mathrm{L}(\Gamma)$ be a propositional language such that its set of primitive propositions consists exactly of the move formulas and payoff formulas for possible moves and payoffs in Γ. Furthermore, there is a knowledge operator K_i for each player i and a common knowledge operator CK. Primitive propositions are well-formed formulas, and if ϕ and ψ are well-formed formulas of $\mathrm{L}(\Gamma)$, so are $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \Rightarrow \psi$, $K_i\phi$, $CK\phi$, and $\phi \hookrightarrow \psi$. The first six kinds of formulas have their usual interpretation "not ϕ", "ϕ and ψ", "ϕ or ψ", "ϕ materially implies ψ", "player i knows ϕ", and "there is common knowledge of ϕ". Formulas of the last kind are interpreted as subjunctive conditionals and will be read as "ϕ conditionally implies ψ". I use the abbreviation $K^n\phi$ defined below to refer to n-th level knowledge of ϕ. Common knowledge of ϕ is the same as n-th level knowledge of ϕ for all natural numbers n. I denotes the set of players.

$$K^1\phi :\Leftrightarrow \bigwedge_{i\in I} K_i\phi$$
$$K^n\phi :\Leftrightarrow \bigwedge_{i\in I} K_iK^{n-1}\phi \qquad \textit{for } n > 1$$

For the epistemic part of the logic, axioms $K0$–$K5$ are valid.

$(K0)$ *all propositional tautologies*
$(K1)$ $K_i\phi \wedge K_i(\phi \Rightarrow \psi) \Rightarrow K_i\psi$
$(K2)$ $K_i\phi \Rightarrow \phi$
$(K3)$ $K_i\phi \Rightarrow K_iK_i\phi$
$(K4)$ $\neg K_i\phi \Rightarrow K_i\neg K_i\phi$
$(K5)$ $CK\phi \Rightarrow \bigwedge_{i\in I} K_i(\phi \wedge CK\phi)$

All of these axioms are widely used in the literature on epistemic foundations of game theory. $K1$ can be interpreted to say that the players have already drawn all logical consequences from their knowledge. $K2$ states that what is known must be true. $K3$ says that the players know what they know and $K4$ that they know what they do not know. $K5$ makes sure that common knowledge indeed implies n-th level knowledge for all natural numbers n.

For the conditional part of the logic, let axioms $C1$–$C6$ determine the properties of conditionals.

$$
\begin{array}{ll}
(C1) & \phi \hookrightarrow \phi \\
(C2) & ((\phi \hookrightarrow \psi_1) \wedge (\phi \hookrightarrow \psi_2)) \Rightarrow (\phi \hookrightarrow (\psi_1 \wedge \psi_2)) \\
(C3) & ((\phi_1 \hookrightarrow \psi) \wedge (\phi_2 \hookrightarrow \psi)) \Rightarrow ((\phi_1 \vee \phi_2) \hookrightarrow \psi) \\
(C4) & ((\phi \hookrightarrow \psi) \wedge (\psi \hookrightarrow \phi)) \Rightarrow ((\phi \hookrightarrow \sigma) \Rightarrow (\psi \hookrightarrow \sigma)) \\
(C5) & ((\phi_1 \hookrightarrow \phi_2) \wedge (\phi_1 \hookrightarrow \psi)) \Rightarrow ((\phi_1 \wedge \phi_2) \hookrightarrow \psi) \\
(C6) & (\phi \hookrightarrow \psi) \Rightarrow (\phi \Rightarrow \psi)
\end{array}
$$

Even though there is no complete agreement about the appropriate properties of subjunctive conditionals in the literature (cf., e.g., Nute, 1984), axioms $C1$–$C6$ appear to be rather uncontroversial. Thus, $C1$ simply says that any proposition ϕ conditionally implies itself, and $C2$ says that if ϕ conditionally implies both ψ_1 and ψ_2, it also conditionally implies their conjunction. Similarly, $C3$ states that if both ϕ_1 and ϕ_2 conditionally imply ψ, then so does their disjunction. $C4$ can be interpreted to mean that if two propositions are conditionally equivalent in the sense of conditionally implying each other, anything conditionally implied by the one is also conditionally implied by the other. $C5$ says that if ϕ_1 conditionally implies both ϕ_2 and ψ, then ϕ_1 and ϕ_2 together also conditionally imply ψ. Finally, $C6$ states that if ϕ conditionally implies ψ and is true, ψ must also be true.

The rules of inference are modus ponens and the following three rules for conditionals CR, for knowledge KR and for common knowledge CKR:

$$
\begin{array}{ll}
(CR) & From\ \phi_1 \Rightarrow \phi_2\ infer\ (\psi \hookrightarrow \phi_1) \Rightarrow (\psi \hookrightarrow \phi_2) \\
(KR) & From\ \phi\ infer\ K_i\phi \\
(CKR) & From\ \phi \Rightarrow \bigwedge_{i \in I} K_i(\phi \wedge \psi)\ infer\ \phi \Rightarrow CK\psi
\end{array}
$$

To give a semantics to this logic, consider conditional epistemic models $(\Omega,\ \mathcal{K}_1, \ldots, \mathcal{K}_n, f, p)$. Ω is a set of states, $\mathcal{K}_i$ is an equivalence relation on Ω, and $p : \Omega \times \{primitive\ propositions\} \rightarrow \{true,\ false\}$ is a valuation function assigning truth values to the primitive propositions for each state. I will refer to $f : \Omega \times \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\Omega)$ as a selection function. For a given state ω and a formula ϕ, $f(\omega, [\phi])$ can intuitively be interpreted as the set of states most similar to ω where ϕ is true. These states are used to determine whether a conditional $\phi \hookrightarrow \psi$ is true at state ω. $[\phi]$ here stands for the set $\{\omega \in \Omega \mid \omega \models \phi\}$, where as usual $\omega \models \phi$ means that ϕ is true at ω. Let restrictions $R1$–$R5$ be imposed on f.

$$
\begin{array}{lll}
(R1) & \omega' \models \phi & if\ \omega' \in f(\omega, [\phi]) \\
(R2) & f(\omega, [\phi_1 \wedge \phi_2]) \subseteq f(\omega, [\phi_1]) & if\ f(\omega, [\phi_1]) \subseteq [\phi_2] \\
(R3) & f(\omega, [\phi_1 \vee \phi_2]) \subseteq f(\omega, [\phi_1]) \cup f(\omega, [\phi_2]) & \\
(R4) & f(\omega, [\phi]) = f(\omega, [\psi]) & if\ f(\omega, [\phi]) \subseteq [\psi]\ and\ f(\omega, [\psi]) \subseteq [\phi] \\
(R5) & \omega \in f(\omega, [\phi]) & if\ \omega \in [\phi]
\end{array}
$$

Truth values of well-formed formulas are defined recursively for a state $\omega \in \Omega$ by the following rules, where $\mathcal{K}$ stands for the transitive closure of $\bigcup_{i \in I} \mathcal{K}_i$:

$$
\begin{array}{ll}
\omega \models \phi & if\ \phi\ is\ primitive\ and\ p(\omega)(\phi) = true \\
\omega \models \neg\phi & if\ not\ \omega \models \phi \\
\omega \models \phi \Rightarrow \psi & if\ not\ \omega \models \phi\ or\ \omega \models \psi \\
\omega \models K_i\phi & if\ \omega' \models \phi\ \forall\omega'\ (\omega, \omega') \in \mathcal{K}_i \\
\omega \models \phi \hookrightarrow \psi & if\ \omega' \models \psi\ \forall\omega' \in f(\omega, [\phi]) \\
\omega \models CK\phi & if\ \omega' \models \phi\ \forall\omega'\ (\omega, \omega') \in \mathcal{K}
\end{array}
$$

It can be shown that the above axiomatization is sound and complete for the class of all conditional epistemic models (for the techniques needed to show this, see, e.g., Halpern, 1998b, and Halpern and Moses, 1992).

3 Description of the Game

In the language $\mathrm{L}(\Gamma)$, the structure of Γ can be described in a straightforward way. Any move to a terminal node should conditionally imply the payoffs associated with this terminal node, and any move leading to a decision node should imply that exactly one of the possible moves at this decision node will be played. Furthermore, any move should imply all moves preceding it in the game tree. Finally, exactly one of the possible moves at the root of the game tree should be played. For the game in Fig. 1, this intuition yields the following formula:

$$
\begin{array}{c}
(U_b \hookrightarrow ([\pi_1 = 4] \wedge [\pi_2 = 2] \wedge U_a)) \wedge (D_b \hookrightarrow ([\pi_1 = 2] \wedge [\pi_2 = 1] \wedge U_a)) \wedge \\
(U_c \hookrightarrow ([\pi_1 = 1] \wedge [\pi_2 = 3] \wedge D_a)) \wedge (D_c \hookrightarrow ([\pi_1 = 3] \wedge [\pi_2 = 4] \wedge D_a)) \wedge \\
(U_a \hookrightarrow ((D_b \vee U_b) \wedge \neg(D_b \wedge U_b))) \wedge (D_a \hookrightarrow ((D_c \vee U_c) \wedge \neg(D_c \wedge U_c))) \wedge \\
(U_a \vee D_a) \wedge \neg(U_a \wedge D_a)
\end{array}
$$

For the general case, some additional terminology is needed. Let F_Γ be the set of all moves that lead to a terminal node, M_Γ the set of all possible moves, $M_\Gamma(v)$ the set of all possible moves at node v, r the root of the game tree, $P_\Gamma(v)$ the set of moves on the path from the root to v, and u_{X_v} the node X_v leads to. Furthermore, $\pi_i(X_v)$ is player i's payoff at the terminal node reached by X_v, and $\otimes$ denotes the "exclusive or", i.e. a truth functional connective defined as follows:

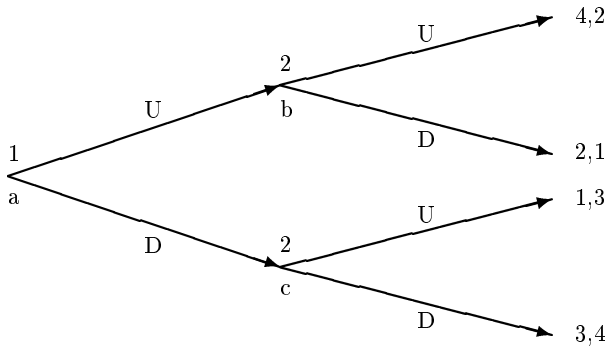

Fig. 1. A generic game of perfect information

$$\bigotimes_{i\in J}\phi_i :\Leftrightarrow \bigvee_{i\in J}(\phi_i \wedge \bigwedge_{j\in J\setminus\{i\}}\neg\phi_j)$$

Now the structure of Γ is described by the following formula S_Γ:

$$\bigwedge_{X_v\in F_\Gamma}(X_v \hookrightarrow \bigwedge_{i\in I}[\pi_i = \pi_i(X_v)] \wedge \bigwedge_{X_s\in P_\Gamma(v)} X_s)\wedge$$
$$\bigwedge_{X_v\in M_\Gamma\setminus F_\Gamma}(X_v \hookrightarrow (\bigwedge_{X_s\in P_\Gamma(v)} X_s \wedge \bigotimes_{X_u\in M_\Gamma(u_{X_v})} X_u)) \wedge \bigotimes_{X_r\in M_\Gamma(r)} X_r$$

4 Two Notions of Rationality

In the literature, two different concepts of rationality can be distinguished which I will call behavioural and habitual rationality. In either case, rationality means that if a player knows that choosing move X_v will result in a higher payoff than choosing move Y_v, he will not choose Y_v. However, with the behavioural concept, this concerns only the actual, observable behaviour of the player. Thus player i's being rational in the behavioural sense allows only to draw conclusions about moves he will not play in the actual world, i.e. at nodes that will actually be reached. Nothing can be said about what would happen at unreached nodes. In contrast to this, the habitual concept allows to draw conclusions about choices at any nodes, independently of whether these will be reached or not. Rationality thus appears as a property of the player which he will keep in any decision situation in which he might find himself. In the literature, this idea of resilient rationality has been objected to as implausible (see, e.g., Rabinowicz, 1998). However, if you think of a player's rationality in terms of how he makes decisions, it seems more natural to assume that he will use the same decision procedure

- his habitual decision procedure - at any node where he might find himself. Consider in particular a situation where the players have initial information about the structure of the game and about the fact that their opponents are rational. Before the start of the game, they try to derive from this information which moves they should choose. In this case a player thinking about how one of his opponents would behave at some node has no reason to doubt that the opponent will use his usual decision procedure there. In fact, as at this point of time he does not yet know which nodes will eventually be reached, treating reached and unreached nodes in an asymmetric way seems to make little sense for the player. Thus if you think of the players' choice of moves as the result of deliberation about how their opponents will play, habitual rationality appears to be the more appropriate concept.

Let me now turn to the formalisation of the two notions. In the language $\mathrm{L}(\Gamma)$, a consequence of behavioural rationality may be expressed by the following formula with $x > y$, where $i(v)$ stands for the player that is to move at v.

$$K_{i(v)}((X_v \hookrightarrow [\pi_{i(v)} = x]) \wedge (Y_v \hookrightarrow [\pi_{i(v)} = y])) \Rightarrow \neg Y_v$$

This formula can be seen as a scheme where the move and payoff formulas can be substituted by other move and payoff formulas (where the first payoff formula always refers to a higher payoff than the second one). The conjunction of the finitely many such substitution instances for all nodes v, all moves at node v, and all possible payoffs gives the behavioural rationality formula, which will be denoted by R^B.

For a formula that says that backward induction is played, write X_v^* for the backward induction move at v. A formula BI can then be defined as follows.

$$BI :\Leftrightarrow X_r^* \wedge \bigwedge_{X_s \in M_\Gamma \setminus F_\Gamma} (X_s \hookrightarrow X^*_{u_{X_s}})$$

Note that BI not only says that the backward induction play will arise, but also that at any unreached node the respective backward induction move would be chosen if it were reached. BI thus captures the choice of backward induction strategies. Now the following negative result can be proven.

Theorem 1. *Common knowledge of behavioural rationality and the structure of the game does not imply backward induction.*

$$\nvdash CK(R^B \wedge S_\Gamma) \Rightarrow BI$$

Proof. To show the unprovability of $CK(R^B \wedge S_\Gamma) \Rightarrow BI$, I will construct a conditional epistemic model containing a state where this formula is false. I will say that a state of a model corresponds to a given play of a given game exactly if it satisfies all and only move formulas X_v and payoff formulas $[\pi_i = x]$ such that X_v belongs to that play and x is player i's payoff if that play obtains. With this terminology, construct a model $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, f, p)$ where Ω contains exactly one state corresponding to each of the four possible plays of the game

in Fig. 1. Let $\mathcal{K}_i = \{(\omega, \omega) | \omega \in \Omega\}$, which is obviously an equivalence relation. In accordance with the above definition of correspondence between states and plays, define $p(\omega)(X_v) = true$ if and only if X_v belongs to the play corresponding to ω, and define $p(\omega)([\pi_i = x]) = true$ if and only if x is player i's payoff if the play corresponding to ω obtains.

Call the state satisfying D_a and D_c ω_1, the one satisfying D_a and U_c ω_2, the one satisfying U_a and D_b ω_3, and the one satisfying U_a and U_b ω_4. Define f as follows:

$$f(\omega_i)([\phi]) = \begin{cases} \{\omega_i\} & if\ \omega_i \models \phi \\ \{\omega_j \mid j = \min\{k \mid \omega_k \in [\phi]\}\} & otherwise \end{cases}$$

A selection function thus defined satisfies restriction $R1$–$R5$ (cf. Clausing, 1999).

Clearly $\omega_1 \not\models BI$. It remains to show $\omega_1 \models CK(R^B \wedge S_\Gamma)$. Because of $\mathcal{K} = \mathcal{K}_i,\ i = 1, 2$, it suffices to show $\omega_1 \models R^B \wedge S_\Gamma$. $\omega_1 \models S_\Gamma$ is easily checked. As ω_1 satisfies only the move formulas D_a and D_c, the only substitution instances of the behavioural rationality scheme that could be false at this state are those with the consequents $\neg D_a$ and $\neg D_c$. However, because of $f(\omega_1, [U_a]) = \{\omega_3\}$ one finds $\omega_1 \models K_1((U_a \hookrightarrow [\pi_1 = u]) \wedge (D_a \hookrightarrow [\pi_1 = d]))$ only for $u = 1$ and $d = 3$. Analogously, one has $\omega_1 \models K_2((U_c \hookrightarrow [\pi_2 = u]) \wedge (D_c \hookrightarrow [\pi_2 = d]))$ only for $u = 3$ and $d = 4$. Consequently, any substitution instance of the behavioural rationality scheme with a false consequent must have a false antecedent, which gives $\omega_1 \models R^B$. □

Note that the proof of theorem 1 does not only show that common knowledge of behavioural rationality and the structure of the game does not imply the choice of backward induction strategies, it also shows that not even the backward induction path is implied.

For a definition of habitual rationality, consider the following formula with $x > y$. Let v be any decision node other than the root and let X_u^v stand for the move leading to v.

$$K_{i(v)}((X_v \hookrightarrow [\pi_{i(v)} = x]) \wedge (Y_v \hookrightarrow [\pi_{i(v)} = y])) \Rightarrow (X_u^v \hookrightarrow \neg Y_v)$$

As demanded above, the conclusion that Y_v will not be played is independent of whether v will actually be reached or not. This formula can again be treated as a scheme. For moves at the root, take the same formula as for behavioural rationality. I will denote the conjunction of the finitely many substitution instances of the habitual rationality scheme for all nodes v, all possible moves at v, and all possible payoffs by R^H. Note that in the presence of S_Γ, R^H implies R^B via $C6$.

Let the length of a decision node v be the maximum number of moves in a play of the subgame beginning at v and denote it by $l(v)$. The length of a game is the length of its root. Define $K^0\phi :\Leftrightarrow \phi$. The following results can now be established.

Theorem 2. *For a game of length m, m-th level knowledge of the structure of the game and m-1-th level knowledge of habitual rationality imply backward induction.*

$$\vdash (K^m S_\Gamma \wedge K^{m-1} R^H) \Rightarrow BI$$

The proof of this theorem makes use of the following lemma. For its formulation, extend the definition of $\pi_i(X_v)$ to non-terminal moves X_v so that it denotes player i's payoff if after X_v, only backward induction moves are played. Define an empty conjunction to be true and let $F(v)$ denote the set of all decision nodes weakly following v (i.e. $v \in F(v)$). As above, X_s^t stands for the move leading to node t and X_t^* is the backward induction move at t.

Lemma 1. *For any node v and any move Y_v, one finds:*

$$\vdash (S_\Gamma \wedge \bigwedge_{t \in F(v) \setminus \{v\}} (X_s^t \hookrightarrow X_t^*)) \Rightarrow (Y_v \hookrightarrow \bigwedge_{i \in I} [\pi_i = \pi_i(Y_v)])$$

Proof (of the lemma). I will use induction on $l(v)$. The base case with $l(v) = 1$ and $F(v) \setminus \{v\} = \emptyset$ is trivial. For the induction step, assume the claim of the lemma has been shown for all nodes w with $l(w) \leq n$ for some natural number $n < m$. Let $l(v) = n + 1$. For $Y_v \in F$, the claim is again trivial. For $Y_v \notin F$, there is a node u with $l(u) \leq n$ to which Y_v leads. Now CR and the induction hypothesis yield:

$$\vdash (S_\Gamma \wedge \bigwedge_{t \in F(v) \setminus \{v\}} (X_s^t \hookrightarrow X_t^*)) \Rightarrow ((X_u^* \hookrightarrow Y_v) \wedge (X_u^* \hookrightarrow \bigwedge_{i \in I} [\pi_i = \pi_i(X_u^*)]))$$

Because $Y_v \hookrightarrow X_u^*$ is one of the conjuncts of the left hand side of this and $\pi_i(X_u^*) = \pi_i(Y_v)$ for all $i \in I$, $C4$ now gives:

$$\vdash (S_\Gamma \wedge \bigwedge_{t \in F(v) \setminus \{v\}} (X_s^t \hookrightarrow X_t^*)) \Rightarrow (Y_v \hookrightarrow \bigwedge_{i \in I} [\pi_i = \pi_i(Y_v)])$$

□

Proof (of the theorem). Let $NBI(v)$ stand for the set of all non backward induction moves at node v. I will again use induction on $l(v)$. For the base case, assume $l(v) = 1$, i.e. all possible moves at v terminate the game. Let Y_v be any move other than the backward induction move. CR yields:

$$\vdash S_\Gamma \Rightarrow ((X_v^* \hookrightarrow [\pi_{i(v)} = \pi_{i(v)}(X_v^*)]) \wedge (Y_v \hookrightarrow [\pi_{i(v)} = \pi_{i(v)}(Y_v)]))$$

KR and $K1$ give:

$$\vdash K^1 S_\Gamma \Rightarrow K_{i(v)}((X_v^* \hookrightarrow [\pi_{i(v)} = \pi_{i(v)}(X_v^*)]) \wedge (Y_v \hookrightarrow [\pi_{i(v)} = \pi_{i(v)}(Y_v)]))$$

If v is the root, i.e. $m = 1$, one finds because of $\pi_{i(v)}(X_v^*) > \pi_{i(v)}(Y_v)$:

$$\vdash (K^1 S_\Gamma \wedge R^H) \Rightarrow \neg Y_v$$

As Y_v was an arbitrary non backward induction move, this gives:

$$\vdash (K^1 S_\Gamma \wedge R^H) \Rightarrow \bigwedge_{Y_v \in NBI(v)} \neg Y_v$$
$$\vdash (K^1 S_\Gamma \wedge R^H) \Rightarrow X_v^*$$

The last step uses $K2$. For $m = 1$, the right hand side is equivalent to BI, and I am done.

If v is not the root, i.e. $m > 1$, there is a move X_u^v leading to v and one finds:

$$\vdash (K^1 S_\Gamma \wedge R^H) \Rightarrow (X_u^v \hookrightarrow \neg Y_v)$$

Again due to the arbitrariness of Y_v, this gives:

$$\vdash (K^1 S_\Gamma \wedge R^H) \Rightarrow \bigwedge_{Y_v \in NBI(v)} (X_u^v \hookrightarrow \neg Y_v)$$

CR and propositional reasoning yield:

$$\vdash (K^1 S_\Gamma \wedge R^H) \Rightarrow (X_u^v \hookrightarrow X_v^*)$$

For the induction step, assume the following has been shown for all decision nodes w with $l(w) \leq n,\ n < m$:

$$\vdash (K^n S_\Gamma \wedge K^{n-1} R^H) \Rightarrow \bigwedge_{t \in F(w)} (X_s^t \hookrightarrow X_t^*)$$

Let $l(v) = n + 1$. KR and $K1$ yield:

$$\vdash (K^{n+1} S_\Gamma \wedge K^n R^H) \Rightarrow K_{i(v)} \Big(\bigwedge_{t \in F(v) \setminus \{v\}} (X_s^t \hookrightarrow X_t^*) \Big)$$

As above, let Y_v be any non backward induction move. Now lemma 1, KR and $K1$ yield:

$$\vdash (K^{n+1} S_\Gamma \wedge K^n R^H) \Rightarrow$$
$$K_{i(v)} ((X_v^* \hookrightarrow [\pi_{i(v)} = \pi_{i(v)}(X_v^*)]) \wedge (Y_v \hookrightarrow [\pi_{i(v)} = \pi_{i(v)}(Y_v)]))$$

If v is the root, i.e. $m = n + 1$, one finds:

$$\vdash (K^{n+1} S_\Gamma \wedge K^n R^H) \Rightarrow \bigwedge_{Y_v \in NBI(v)} \neg Y_v$$
$$\vdash (K^{n+1} S_\Gamma \wedge K^n R^H) \Rightarrow \Big(\bigwedge_{t \in F(v) \setminus v} (X_s^t \hookrightarrow X_t^*) \wedge X_v^* \Big)$$

Again, the right hand side is equivalent to BI.
If $m > n+1$, there is a move X_u^v and one finds:

$$\vdash (K^{n+1}S_\Gamma \wedge K^n R^H) \Rightarrow \bigwedge_{Y_v \in NBI(M_\Gamma(v))} (X_u^v \hookrightarrow \neg Y_v)$$

$$\vdash (K^{n+1}S_\Gamma \wedge K^n R^H) \Rightarrow \bigwedge_{t \in F(v)} (X_s^t \hookrightarrow X_t^*)$$

□

Note that the one-level difference between the necessary knowledge about the structure of the game and about rationality in the statement of theorem 2 appears quite natural if you consider a game of length one, i.e. a one-person decision problem. To reach an optimal decision, the person certainly has to know something about the structure of his problem, but not about anyone's rationality.

The consistency of the antecedent of theorem 2 is an immediate corollary of the following result.

Theorem 3. *Common knowledge of habitual rationality and the structure of the game is a consistent assumption.*

Proof. As in the proof of theorem 1, construct a model $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, f, p)$ where Ω contains for every possible play of the game exactly one state corresponding to it and define p accordingly. Let $\mathcal{K}_i = \{(\omega,\omega)|\omega \in \Omega\}$.

For the definition of the state selection function, assign natural numbers to the plays of the game by the following recursive procedure. Assign number 1 to the backward induction play. If numbers up to n have been assigned, take the last node reached by play number n where there are moves which do not belong to any of the first n plays. Call this node the departing node for play $n+1$. From the moves at the departing node that do not belong to any of the first n plays, choose the one which gives the player making it the highest payoff under the assumption that at all succeeding nodes, only backward induction moves are played. This choice is unique as the game is generic by assumption. Now assign number $n+1$ to the play which contains this chosen move and where all moves following it are backward induction moves (i.e. at all nodes reached after the departing node, backward induction moves are played). It is easy to check that this procedure assigns a unique number to all plays.

Now assign to each state the number of its corresponding play and define f as in the proof of theorem 1.

I now show $\omega_1 \models CK(R^H \wedge S_\Gamma)$. For this it is clearly sufficient to show $\omega_1 \models R^H \wedge S_\Gamma$. It is not hard to see that S_Γ is true at ω_1. Furthermore, one can check that for any move X_v, the unique world contained in $f(\omega_1, [X_v])$ corresponds to the play where from X_v onwards, only backward induction moves are chosen. Consequently, $\omega_k \in f(\omega_1, [X_v])$ implies $\omega_k \models [\pi_{i(v)} = \pi_{i(v)}(X_v)]$ and one thus finds $\omega_1 \models X_v \wedge [\pi_{i(v)} = \pi_{i(v)}(X_v)]$ and $\omega_1 \models\!\!\not\; X_v \hookrightarrow [\pi_{i(v)} = x]$ for any $x \neq \pi_{i(v)}(X_v)$. Therefore $\omega_1 \models K_{i(v)}(X_v \hookrightarrow [\pi_{i(v)} = y])$ if and only if

$y = \pi_{i(v)}(X_v)$. Furthermore, if Y_w is a non backward induction move at the node to which X_v leads, it must be $\omega_k \models \neg Y_w$ and thus $\omega_1 \models X_v \hookrightarrow \neg Y_w$. Because of $\omega_1 \models \neg Y_r$ for any non backward induction move at the root, one finds that the consequent of any instance of habitual rationality containing a negated non backward induction move is satisfied by ω_1. For instances containing negated backward induction moves X_v^* in their consequents, the antecedent cannot be satisfied at ω_1 because of $\pi_{i(v)}(X_v^*) > \pi_{i(v)}(Y_v)$ for any alternative move Y_v. This means $\omega_1 \models R^H$. □

As behavioural rationality is implied by habitual rationality in the presence of S_Γ, the consistency of common knowledge of behavioural rationality and the structure of the game is a corollary to theorem 3.

5 Belief instead of Knowledge

The axioms $K1$–$K5$ describe different properties of knowledge. There is an extensive literature on the appropriateness of these axioms some of which are rather controversial. Thus Binmore and Shin (1992) argue that $K2$ should be abandoned in an analysis of knowledge in games. Stalnaker (1996) writes that $K4$ is not reasonable as a property of knowledge. Therefore it is of some interest that versions of the above results remain valid if the axioms $K2$–$K4$ are abandoned. As usual in the absence of $K2$, I will talk of belief instead of knowledge and write the operators as B_i, B^n and CB. I will say there is true common belief in ϕ if $CB\phi \wedge \phi$ is true and accordingly for true n-th level belief. Consider a logic with the axioms $K0$, $K1$, $K5$, and $C1$–$C6$ and the rules of inference modus ponens, CR, KR, and CKR. It can be shown that this axiomatization is sound and complete for the class of all conditional doxastic models $(\Omega, \mathcal{B}_1, \ldots, \mathcal{B}_n, f, p)$, where Ω, f and p are as for conditional epistemic models and the $\mathcal{B}_i$ are arbitrary binary relations on Ω.

It follows immediately from the definition of $(\Omega, \mathcal{B}_1, \ldots, \mathcal{B}_n, f, p)$ that the models constructed in the proofs of theorems 1 and 3 are also conditional doxastic models, which shows that true common belief in the structure of the game and behavioural rationality as well as true common belief in the structure of the game and habitual rationality are consistent assumptions. It furthermore shows that the former does not imply backward induction. Finally the proof of theorem 2 can be copied with minor adaptations to prove the following.

Theorem 4. *For a game of length m, true m-th level belief in the structure of the game and true m-1-th level belief in habitual rationality imply backward induction.*

$$\vdash (B^m S_\Gamma \wedge B^{m-1} R^H \wedge S_\Gamma \wedge R^H) \Rightarrow BI$$

6 Related Literature

The idea of representing the structure of a game by a logical formula goes back to Bonanno (1991). He, however, uses a purely propositional logic, which does

not allow to capture the conditional aspects of a game as represented by the game tree.

In Samet (1996), sufficient conditions for backward induction play are presented that are based on the use of a hypothetical knowledge operator. Halpern (1998a) has shown that this operator can also be expressed with the help of knowledge operators and conditionals. These conditionals, however, are "epistemic" in the sense that they refer to the players' states of mind rather than to the real world. Consequently, like knowledge operators they are indexed by the players. In my logic, on the contrary, conditionality is seen as an objective feature of the real world. Furthermore, like most of the literature on epistemic foundations for solution concepts, Samet assumes that players know which moves they will take. In Rabinowicz's (1998) terminology, he employs an at-choice perspective. I do not make this assumption and thus employ a pre-choice perspective. This is certainly in line with the story told above about players trying to decide which moves to choose by conscious deliberation based on some initial information. In particular, the pre-choice perspective does not a priori exclude the possibility that deliberation will not lead to a uniquely determined decision, as may well be the case. In this connection, a word on the point of time at which the players have the knowledge that is modelled in my conditional epistemic logic is in order. I take it to be before the game is played and after the players have completed their deliberations about the game. The latter is clearly what is captured by axiom $K1$. However, the point of time when a player has completed his deliberation may well be before he has reached a decision about which moves to play if deliberation does not yield a unique choice.

The term habitual rationality is taken from Aumann (1995), who calls the players in his model habitual payoff maximizers. He derives sufficient conditions for backward induction with the help of a conditional payoff function which captures the conditional aspects of the game and the players' strategies in an implicit way. In my approach, these aspects can be treated explicitly as it is moves rather than strategies which are taken as primitives.

References

Aumann, R.: Backward Induction and Common Knowledge of Rationality. Games and Economic Behavior **8** (1995) 6–19

Binmore, K., Shin, H.: Algorithmic knowledge and game theory. In: Bicchieri and Dalla Chiara (eds.), Knowledge, Belief, and Strategic Interaction, Cambridge MA (1992) 141–154

Bonanno, G.: The Logic of Rational Play in Games of Perfect Information. Economics and Philosophy **7** (1991) 37–65

Clausing, T.: The Logical Modelling of Reasoning Processes in Games. Doctoral thesis, Leipzig (1999)

Dekel, E., Gul, F.: Rationality and common knowledge in game theory. In: Kreps and Wallis (eds.), Advances in Economics and Econometrics: Theory and Application, Vol. I, Cambridge (1996), 87–172

Halpern, J.: Set-Theoretic Completeness for Epistemic and Conditional Logic. Mimeo (1998a)

Halpern, J.: Hypothetical Knowledge and Counterfactual Reasoning. Mimeo (1998b)
Halpern, J., Moses, Y.: A guide to completeness and complexity for modal logics of knowledge and belief. Artificial Intelligence **54** (1992) 311–379
Nute, D.: Conditional Logic. In: Gabbay and Guenthner (eds.), Handbook of Philosophical Logic, Vol. II, Dordrecht (1984), 387–439
Rabinowicz, W.: Grappling with the Centipede: Defence of Backward Induction for BI-Terminating Games. Economics and Philosophy **14** (1998) 95–126
Samet, D.: Hypothetical Knowledge and Games with Perfect Information. Games and Economic Behavior **17** (1996) 230–251
Stalnaker, R.: Knowledge, Belief and Counterfactual Reasoning in Games. Economics and Philosophy **12** (1996) 133–163