The Backward Induction Solution to the Centipede Game*

Graciela Rodríguez Mariné
University of California, Los Angeles
Department of Economics
November, 1995

Abstract

In extensive form games of perfect information, where all play could potentially be observed, the backward induction algorithm yields strategy profiles whose actions are best responses at every possible subgame. To find these actions, players must deliberate about the outcomes of their choices at every node they may be called to play, based upon their mutual knowledge of rationality. However, there are in general nodes that will not reached under equilibrium, and in these situations, players must *hypothesize* about the truth of counterfactuals asserting *what would have happened had a deviation occurred*. The paper conjectures that deviations *may* confer *information* relevant for future play and therefore have a *causal consequence* upon contingent play. A proper foundation for the backward induction solution requires therefore, the formalization of strategies as contingent constructions *as well as* a theory of counterfactuals to support the truth condition of these conditionals. The paper considers Lewis´ and Bennett's criteria to assert the truth of counterfactual conditionals and conjectures that these approaches lead to different ways of thinking about deviations. According to our interpretation of Lewis´approach and in our version of the centipede game, *common knowledge* of *rationality* –as it is defined in the paper- leads to the backward induction outcome. According to our interpretation of Bennett´s approach, backward induction can be supported only if the players have the necessary amount of ignorance, which depends on the number of nodes of the game.

---

*This paper is a revised version of the first chapter of the doctoral dissertation submitted to UCLA in November, 1995.

1 Introduction

Reasoning about the outcomes of alternative actions is a crucial constituent of any decision. A player can not rationally *choose* a strategy if he can not *assert* what *would have happened* otherwise. In particular, the play of a given equilibrium by a player is justified in terms of his rationality, if he either knows or believes that *had he deviated he would have not been better off.* In other words, conjectures about the occurrence of events that are *not expected under equilibrium* not only support or justify the choice of a strategy, but also assure that it is not profitable to deviate**.**

Consider an extensive game of perfect information. If players deliberate about their decisions at every subgame and therefore optimize in each possible scenario on- and off-the-equilibrium- path, then, not only unilateral deviations will be unprofitable (a requisite that every Nash equilibrium satisfies) but also deviations by more than one player.

This is the idea upon which the backward induction argument is based. Yet the problem with the algorithm, as it is typically presented, is that the corresponding counterfactual reasoning is not analyzed as such. Deviations are devoid of meaning and hence, are not supposed to confer any *information* to the players regarding the rationality of the deviator. This means that they can not have *consequences* upon contingent play, which ultimately depends on the maximizing choice at the last node.

Our main premise is that, in order to obtain a proper foundation for the backward induction algorithm, players need an appropriate framework to assert the truth condition of the conditionals involved in deliberation off-the-equilibrium path. As it is extensively acknowledged in the literature, the outcomes of these thought experiments will depend not only upon this framework, but also upon the knowledge and beliefs of the players regarding the game and their mutual rationality.

In the literature of non cooperative extensive form games of perfect information, the centipede game is one whose backward induction solution still motivates a considerable amount of disagreement concerning its logical foundations. The solution is also considered counterintuitive or puzzling and does not perform well in experimental studies [16]. Two issues sustain the theoretical controversy. On the one hand, there is the question of how to give meaning to the assumption of rationality in the context of counterfactual reasoning and on the other, assuming that this is possible, how to derive the backward induction outcome from this supposition.

With respect to the first issue, Reny [17] asserts that common knowledge of rationality is not attainable in games exhibiting the properties of the centipede game. After observing a deviation in a centipede game with three or more nodes, there cannot be common knowledge that the players are maximizers. On the other hand, Binmore [7] asserts that the

irrationality of a player who deviates in the centipede game is an open matter, because it is not clear what the opponent should deduce about the rationality and further play of the deviator.

To concentrate on the second question, let us assume that it is possible for the players to have common knowledge of rationality. The issue of how to derive the backward induction outcome when hypothetical thinking is present, is also a matter of controversy.

Aumann [2] proves that in games of perfect information, common knowledge of rationality leads to the backward induction equilibrium. On the other hand, Binmore [7] claims that rational players would not necessarily use the strategies resulting from this algorithm. He supports the equilibrium where the first player plays his backward induction strategy and the second mixes between leaving and taking the money. In [6], he proposes to enlarge the model by introducing an infinite set of players, so that the presence of irrational players, who exist with probability zero, is not ruled out altogether. Bicchieri ([4]&[5]) proves that, under the assumption of common knowledge of rationality, there is a lower and an upper bound of mutual knowledge that can support the backward induction outcome. The lower bound involves a level of mutual knowledge for the root player equal to the number of nodes in the equilibrium path minus one. Samet [18] proves within his framework, that common hypothesis of rationality at each node implies backward induction and that, for each node off-the-equilibrium path, there is common hypothesis that if that node were to be reached then it would be the case that not all players are rational.

The purpose of this paper is to test the internal consistency of the backward induction algorithm by presenting a formalization capable of incorporating counterfactual reasoning at nodes off-the-equilibrium-path. The aim is to find sufficient conditions, regarding players' knowledge and beliefs, capable of yielding the truth of the supporting counterfactuals.

The paper considers two criteria to determine the truth of counterfactual conditionals, based upon the theories of counterfactuals developed by David Lewis [14] and Jonathan Bennett [3] respectively. Under our interpretation of Lewis' approach and the assumption of common knowledge of rationality, (as it will be defined below) the backward induction outcome can be obtained. The reason is that players are not necessarily led to reject their beliefs concerning the rationality of their opponents at *other* counterfactual scenarios where they might have a chance to play again. Under our interpretation of Bennett's approach and the assumption of common knowledge of rationality, the theory becomes inconsistent. This result is similar in spirit to the one in Bicchieri [5] although it is obtained under different conditions. Unless the amount of mutual knowledge of the root player is reduced to a level equal to the number of nodes in the equilibrium path minus one, backward induction can not be supported. Relaxing the assumption of common *knowledge* or rationality in favor of common *belief* implies that there may be scenarios compatible with backward induction

where no inconsistency obtains although common belief in rationality needs to be dropped in these situations. This result resembles one of the outcomes in Samet [18].

The organization of this paper is as follows. The first section explains the nature of counterfactuals and analyses their role in strategic situations. The second, presents the framework and formalization of the backward induction solution in terms of counterfactual reasoning. The third section incorporates the two mentioned approaches to establish the truth of these counterfactuals and analyzes the conditions, in terms of different levels of mutual knowledge and belief, under which the backward induction outcome obtains. To conclude, the fourth presents an overall evaluation of the results, in perspective with their philosophical justifications and implications.

1.1 Counterfactual conditionals

A counterfactual or a subjunctive conditional is an implication of the following form:

Had P happened then Q would have happened.

The counterfactual connective will be denoted by " $\Box\!\!\rightarrow$ " and the previous subjunctive conditional will be denoted by "P $\Box\!\!\rightarrow$Q", where "P" and "Q" are two propositions defined within some language L[1]. The difference between a counterfactual and an indicative conditional represented by *"If P then Q"* is that P is necessarily false in the case of a counterfactual.

Truth functional analysis establishes that "if P then Q" is true in the following circumstance: Q is true or P is false. If this approach were to be followed in the case of counterfactual conditionals we would be left with no clear result; any conditional with a false antecedent would be true regardless of the truth condition of the consequent. Nevertheless, Stalnaker [20] observes that "the falsity of the antecedent is never sufficient reason to affirm a conditional, even an indicative conditional." Conditionals, no matter whether indicative or subjunctive, establish a connection or function between propositions and this connection is not necessarily represented by the truth functional analysis. The truth functional analysis only deals with the truth conditions of the propositions in isolation yet the conditional alludes to some *connection* or *function* between the propositions.

Within purely logical or mathematical systems the connection between propositions is ruled by a set of axioms. In this case, truth functional analysis is sufficient. However, when conditionals refer to other types of frameworks this criteria is not sufficient. Consider for instance the following conditional: "If John studies for the test, he will pass the exam." Would we try to assert the truth of this conditional by answering whether it is true that John

[1] The expressions: *propositions*, *predicates*, *sentences* or *formulas* will be used indistinctively from now on.

will study and whether it is true that he will pass the exam? The answer is clearly negative. We will say that the conditional is true only if we can support the opinion that studying is enough to pass an exam. Were we to consider that luck is what matters, then it could be true that John studied and passed the exam, but actually did so as a consequence of being lucky.

Counterfactual conditionals are similar to indicative conditionals in this respect. Imagine John did not study and he did not pass the exam. We could say *"had John studied he would have passed the exam"*. Again, consider a purely truth functional analysis. John did not study. Therefore, the antecedent is false and the subjunctive conditional is true regardless of whether he passed the exam. Is this enough to solve the previous counterfactual? Obviously, not. In order to do so, we need to have a hypothesis of how studying could have affected passing the exam. As in the case of indicative conditionals, we need to test whether the *connection,* counterfactual or not, exists.

One approach to the task of solving counterfactuals starts with the premise that the issue of how to assert the truth of a counterfactual is basically the question of how to *inductively* project a predicate (see Goodman [10]). This is a *principle-oriented* criteria because it stresses the existence of a principle that links the predicates that form part of the conditional. Although counterfactuals deal with events that *have not happened* and therefore can not be solved by means of empirical tests, we can construct a criteria based on some observed regularity that represents the connection between the antecedent and the consequent. For instance, a player that decided to play an equilibrium strategy cannot test what would have happened otherwise, because he is not going to deviate. He needs a hypothesis concerning the repercussions of his deviation and this hypothesis cannot be brought about by a test within *this* game. Players may be able to form a hypothesis based on *previous experience* with the same game or players. However, if they decide to play the equilibrium, that is because the "otherwise-hypothesis" has a definite answer[2]. In other words, players cannot run a test while they play the game to discover something they should have known in order to decide *a priori* how to play. When this answer cannot be established players are left with no rational choice. Given that counterfactuals cannot be handled by experimentation or logical manipulation, there is a need for a set of principles to characterize the conditions under which the corresponding predicate can be projected. In the first example, the predicate is "students that study pass exams". To say that *"had John studied he would have passed the exam"* is true, is to assert that the predicate "students that study pass exams" can be extended from a sample to an unobserved case which is John's case.

---

[2] This includes their assigning probability values or ranges when decisions are modeled in uncertain environments.

This approach is not very powerful when we can not identify a principle or predicate to project, when we don't have enough information, or our sample of past predictions is not good enough to trust projections. Consider the counterfactuals involved in game theoretical reasoning. The previous approach would be useful if we thought of behavior in games as determined by a human *disposition*. In this case we would assume that players' behavior is intrinsically ruled by a principle. Players within a game may never fully characterize this principle but at least in certain environments they may be able to construct a well entrenched hypothesis, given their sample of observations. However, this does not apply to games which are not played oft enough for the players to learn something about the behavior of their opponents.

The literature in games has developed a consensus regarding the issue that rational choices are not rational because they are chosen by rational players. In general it is asserted that a person is rational if he chooses rationally (see Binmore [6]&[7]). Leaving this matter aside, we are going to introduce an alternative framework to assert the truth of counterfactuals that seems to be more compatible with this last concept of rationality. This is the approach to counterfactuals in terms of *possible worlds*.

Within the possible-worlds semantics (see Stalnaker [20]) the truth of a counterfactual does not necessarily depend on the existence of a principle or law. To evaluate whether $P \,\Box\!\!\rightarrow Q$ is true one has to realize the following thought experiment: "add the antecedent (hypothetically) to your stock of knowledge (or beliefs), and then consider whether or not the consequent is true" (Stalnaker [20]). When there is a principle or a connection involved, then it should be part of the beliefs that we should hold and we should consider as hypothetically true any consequence that, by this principle, follows from the antecedent. When no connection is suspected or believed, one should analyze the counterfactual in terms of the beliefs in the corresponding propositions, and the relevant issue is whether or not the counterfactual antecedent and consequent can be believed to hold at the same time. Following this approach, which is similar in spirit to Frank Ramsey's test for evaluating the acceptability of hypothetical statements, Stalnaker [20] and Lewis ([14]&[15]) have suggested two closely related theories of counterfactuals (see Harper [11]).

When we believe that the antecedent is false (for instance, when the antecedent entails a deviation by some player) the thought experiment or world, within which the antecedent is true, may not result from the mere addition of the antecedent to the stock of beliefs without resulting in a contradiction. Therefore, the beliefs that contradict the antecedent should be deleted or revised. The problem is that there *is not be a unique way to do so*. A deviation may imply at least one of the following things: i) the deviator is simply irrational either in terms of his reasoning capacities or formation of beliefs, ii) he is rational in terms of his reasoning capacities but he just made a mistake in the implementation of his choice iii) he

did it on purpose, due to the lack of knowledge about his opponents' knowledge or iv) as in iii) but due to the lack of knowledge concerning either the structure of the game or his opponents' rationality.

There is no way to avoid the multiplicity of possible explanations and the issue is that whatever the players believe, it should be commonly held for the equilibrium outcome to be consistent.

Possible world theories offer a framework to evaluate which of the possible explanations should or could be chosen. A possible P-world is an epistemological entity, a state of mind of a player, represented by his knowledge and belief, in which proposition P is true. For instance, the previous four explanations represent possible worlds in which a deviation is believed to have occurred. They are all *deviation-compatible scenarios*. Possible world theories assert, roughly speaking, that in order to evaluate the truth of a counterfactual representing a deviation we need a *criterion* to select which of the above deviation-worlds is the most *plausible*. In the case of game theory, this criterion requires a behavioral assumption that in general is represented by the concept of rationality. We need to find the deviation-world (there could be more than one) that contains the minimal departure from the equilibrium world and evaluate, in terms of players' rationality, which consequent or response, holds *in that closest world*. The equilibrium world will be defined as the *actual* world and we will assume that in *this* world, players are rational (in a suitably defined way) and have some degree of mutual knowledge in their rationality.

1.2 Counterfactuals in Game Theory

Consider the following example that closely resembles off-the-equilibrium path reasoning:

> John is looking down the street standing at the top of the Empire State Building. As he starts walking down the stairs he says to himself: "Hmm, had I jumped off I would have killed myself..."
>
> A very close friend of his is asked later on whether he thinks it is true that "had John jumped off the Empire State building he would have killed himself".
>
> Well, he says, I know John very well; he is a rational person. He would have not jumped off hadn't there been a safety net underneath... I hold that counterfactual is false...[3]

[3] This example is discussed in Jackson [13] and Bennett [3].

Rationality in strategic contexts is a complex phenomenon. There is on the one hand the rationality that alludes to players' capacity to optimize given their knowledge and beliefs and on the other their rationality in terms of belief formation. However, there is a further issue that is particularly critical in games where actions can be observed. Players do not only need to *decide* but to *act upon* their decisions. Moreover, given the fact that actions are *observed*, actual performances will *confer some information* to the other players and therefore may have an impact on their *decisions* about how to further play the game. If a deviation is understood as some non systematical imperfection in the mapping from decisions to actions, then the assumption concerning the rationality in reasoning and belief formation of the deviator does not need to be updated. When this is ruled out, some intentionality must be assumed. When John's friend is asked about the truth of the counterfactual that had John jumping from the top of the building, he is assuming that nothing can go wrong with John's capability to perform what he wants and that therefore, a world in which John jumps, is a world in which a safety net *needs* to exist. There are two issues here. On the one hand, it is reasonable to assume that *in the actual world* John can fully control his capability of not falling in an unintended way, yet this capacity may be deleted in the *hypothetical* world in which he jumps. This relaxation can be considered as a thought experiment that is, the envisagement of a hypothetical world in which the only different fact with respect to the actual world is that John jumps and where no further changes (neither psychological nor physical) interfere with the outcome of the fall. The crucial and troublesome issue in game theory is to establish whether a deviation could imply further deviations by the *same* player. Are these counterfactual worlds correlated?

Another issue is to define which parameters or features of the world we are allowed to change when deliberating about a deviation. Counterfactuals are acknowledged to be *context dependent* and subject to *incomplete specification*. John's friend may know that in the actual world, the one in which John did not jump, there was no safety net. However, in the hypothetical scenario in which John jumps, his friend's willingness to keep full rationality (absence of wrong performances) obliges him to introduce a net. Which similarity with the real world should be preserved? That concerning the safety net or that which assumes that nothing can go wrong? Assume we think that John is rational because he does not typically jump from the top of skyscrapers. This is his *decision*. However, had he either *decided* or *done* otherwise *in that case, where there was no safety net,* he would have died. We would assert that the counterfactual under analysis is true because, although John did not choose to jump he could have done so, and had he jumped off in a world in which the only difference with the actual is John's decision or performance, then he would have killed himself. Is this reasoning the only possible one? It is obviously not. His friend does not seem to think this way.

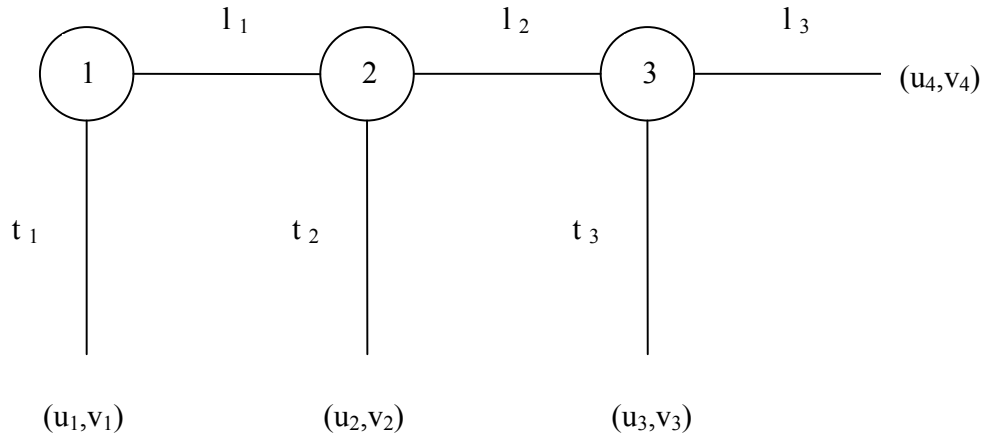
Following the parallel with game theory, consider a case such that if John jumps then his friend will face the decision of whether to jump or not from the same building. Now his reasoning will lead him to the conclusion that jumping must be harmless if John jumps since there must be a safety net at the bottom. If the utility he derives from reaching the floor alive after jumping is higher than the one he gets by not jumping and if he is rational, in the sense of optimizing upon beliefs, then he should *contingently* jump as well! Assume now that the friend's decision should be made before John is actually at the top of the building. Will John's friend jump *contingent* on John's jumping?

In a world in which John jumps, his friend gets some information that makes him change his decision (if we assume he would have not jumped in the absence of a net). However, John's friend could have updated his stock of beliefs to attribute the hypothetical occurrence of the jump to some unexplainable reason but kept the absence of a net, which he believes is a fact in the *actual* world *where he has to decide* whether to jump or not.

2. The backward induction solution to the centipede game

2.1 The centipede game

Consider the following version of the Centipede game: there are two players, called them 1 and 2 respectively. Player 1 starts the game by deciding whether to take a pile of money that lies on the table. If he takes it the game ends and he gets a payoff (or utility value) equal to u_1 whereas his opponent gets v_1. If he leaves the money then player 2 has to decide upon the same type of actions; that is, between taking or leaving the money. Again if she takes it she gets a payoff of v_2 whereas player 1 gets u_2. If player 2 leaves the money then player 1 has the final move. It he takes, player 1 and player 2 get respectively u_3 and v_3. Otherwise, they get payoffs equal to u_4 and v_4 respectively.



The pair of letters between parenthesis at the termination points represent the players' payoffs and they are such that $u_1 > u_2$, $u_3 > u_4$ and $v_2 > v_3$.

The numbers inside the circles represent the labels of the nodes.

t_n stands for taking the money at node n, n=1,2,3.

l_n stands for leaving the money at node n, n=1,2,3.

The backward induction solution to this game has every player taking the money at each node, that is, playing "t_n", for n=1,2,3 whether -on- or -off-the-equilibrium path. The argument briefly says that if player 1 gives player 2 the chance to play he would take the money, for he would expect the first player to do so at the last node. Knowing this, player 1 decides to take the money at the first node.

The controversial issue is that equilibrium play is based upon beliefs at nodes off-the-equilibrium path that do not properly consider how the information which *would be available* at each stage is handled. In the counterfactual hypothesis that the second node is reached, the players are supposed to ignore that something counter to full rationality ought to have occurred, namely, that l_1 has been played. The irrational nature of this play crucially depends on player 1's expectation about the behavior of player 2 at the next node which in turn, depends on player 2's expectation about player 1's further play. Yet, these beliefs are not updated. The relevant information to decide how to play is not what *has been played*, but what it is *expected to be played*. The exception is the last node where the decision depends upon the comparison of payoffs that the player can obtain with certainty. Once some behavioral assumption is introduced, the action to be played at the last node will be determined, given that there are no ties in this game, and this backtracking reasoning will yield a sequence of choices independent of deviations. The question raised in the literature concerns this behavioral assumption at the last node. Why should player 2 expect player 1 to take the money at the third node if he already deviated?

In the past years an agreement concerning the role of counterfactual scenarios has emerged within the literature (see Binmore [7], Bicchieri [5], Samet [18]). Also Aumann [2] who proves that common knowledge of rationality implies backward induction asserts that "substantive conditionals are not part of the formal apparatus, but they are important in interpreting four key concepts"...":strategy, conditional payoff, rationality at a vertex, and rationality" ([2] p.17). He asserts that the "if ...then" clauses involved in equilibrium are not material conditionals but substantive conditionals.[4]

2.2 Definitions and notation

Our version of the centipede game can be represented by:

(1) A finite set of players' labels I, $I=\{i,\ i=1,2\}$.

(2) A finite tree with an order of moves. The set of nodes' labels for players 1 and 2 is denoted respectively by N_1 and N_2 and defined as $N_1 = \{1,3\}$; $N_2 = \{2\}$; The labels represent the order in which players move. The set of all nodes' labels is $N=\{n,\ n=1,2,3\}= N_1 \cup N_2$. $N \subset \mathbf{N}$ (set of natural numbers)

Let Z be the set of terminal nodes' labels. $Z=\{z_1,z_2,z_3,z_4\}$. For each $z \in Z$ there is a unique path leading to it from the initial node. The path leading to the terminal node z is indicated by P(z). Therefore we have:

$P(z_1)=(t_1);P(z_2)=(l_1t_2),P(z_3)=(l_1l_2t_3),P(z_4)=(l_1l_2l_3)$.

(3) A finite set of actions for each player available at each node:

$A_{1n} = \{ a_{1n} ,\ a_{1n} = t_n, l_n \}$ $\quad n=1,3$;

$A_{2n} = \{ a_{2n} ,\ a_{2n} = t_n, l_n \}$ $\quad n=2$;

$A_n = \{ a_n ,\ a_n = t_n, l_n \}$ set of actions available at node n (n=1,2,3).

(4) A public story (h^n) of the game at node n. It consists in the sequence of actions leading to node n from the initial node.[5] In addition let h^{n+1} include the action taken at node n:

$h^{n+1}=\{a_1,.. a_n \}$ $\quad a_n \in A_n$; $n=1,2,3$.

Given that this is a game of perfect information, h^n represents players' knowledge about the past play which leads to node n. Moreover, the set that represents the players' knowledge about the node at which they have to move is a singleton. By definition ($h^1=\varnothing$).

Let H be the set of all terminal histories. Therefore $H=\{P(z_1);P(z_2),P(z_3),P(z_4)\}$

---

[4] He acknowledges that the term "substantive" has been coined by economists only. A substantive conditional is a non material conditional and within his terminology a *counterfactual* is a *substantive conditional* with a *false* antecedent.

[5] This sequence is unique in extensive form games with perfect information.

Let us define $P(z_1) \equiv h^{z_1}$,$P(z_2) \equiv h^{z_2}$;$P(z_3) \equiv h^{z_3}$;$P(z_4) \equiv h^{z_4}$.

(5) A strategy for player i, (i=1,2) is defined as a set of maps. Given some previous history of play, each map assigns to every possible node, at which player i might find himself, an action from the set of feasible actions at that node.

$s_i : N_i \rightarrow A_{in}$; $A_{in} \subset A_i$,$n \in N_i$ i=1,2 ;

The sets of strategies for players 1 and 2 respectively are:

S_1={ t_1t_3 , t_1l_3 , l_1t_3 , l_1l_3 };

S_2={t_2, l_2 }

A strategy profile 's' is a list of strategies one for each player: $s=(s_i)_{i\in I}$

(6) Players' payoffs functions assign to each possible terminal history of the game a real number. U_i :H $\rightarrow$ **R** i=1,2 .

(7) An information structure for each player (also called the player's state of mind) describing the player's knowledge, beliefs and hypotheses.

In order to define these epistemic operators, we need to specify the language within which the framework is defined. This language is constructed upon two types of primitive propositions, or formulas: the ones denoting the play of an action by some player at some node and the ones reflecting the fact that some node has been reached.

These primitive propositions or formulas will be denoted by:

"n", which should be read as "node n is reached" (n=1,2,3)

"a_{in}" , which should be read as "action 'a' is played by player 'i' at node 'n' "

"s_i" , which should be read as " strategy 's' is played by player 'i' ".

Propositions will be generically denoted by P and Q.

The set of primitive formulas is enlarged in the following way:

(*i*) Atomic formulas or primitive predicates (as they have been defined above) are formulas;

(*ii*) if *p* is a formula, then so is "~*p*";

(*iii*) if *p* and *q* are formulas, then so are "(*p*&*q*)" "(*p*v*q*)" and "(*p* $\Box\rightarrow$*q*)";[6]

In addition, the set of primitive formulas is enlarged by the introduction of the following epistemic and doxastic operators:

"K_i" : "i knows that"

"B_i" : "i believes that"

"P_i" : "it is possible, for all that i knows, that"

"C_i" : "it is compatible with everything i knows, that"

---

[6] Notice that within this framework material implications can be expressed in terms of "~" and "&". This is not the case for the counterfactual connective because its truth does not depend on the truth value of its components.

"~*p*" does not refer to the mere result of prefixing "not" to *p*. It refers rather to the corresponding negative sentence, often referred to as the contradictory of *p*.

i is a free individual symbol, that is, it denotes the agent named 'i' and *p* is an arbitrary sentence or predicate.

The last condition to complete the description of our language is:

(*iv*) if *p* is a formula and *i* a free individual symbol (which can take only names of persons as their substitution-values), then $"K_i"$, $"P_i"$, $"B_i"$, and $"C_i"$ are formulas. In each case, *p* is said to be the *scope* of the epistemic operator in question.

2.3 Knowledge and Belief

The study of the concepts of *knowledge* and *belief* together with their uses requires the consideration of a broad set of disciplines due to the complexity that the corresponding phenomena displays. There is on the one hand the obvious semantic and syntactic facets, and on the other, the psychoanalytical one.

In the present essay, we are going to adopt an extremely narrow view of these phenomena. A player *knows* something iff he is actively aware of such a state and has the conviction that there is no need to collect further evidence to support his claim of knowledge. Under this assumption, if it is consistent to utter that *"for all I know it is possible that p is the case"*, then it must be possible for *p* to turn out to be true without invalidating *the knowledge* I claim to have. If somebody claims to know that a certain proposition is true, then the corresponding proposition is true. We rule out the possibility of somebody forgetting something he knew and restrict the environment within which claims of knowledge are considered, to situations in which information does not change. When a new piece of information is acquired, a new instance starts from the epistemological point of view. Moreover, agent's knowledge is supposed to contain not only the primitive notions they are capable to assert they know but also all the logical implications of those sentences.

Although we may show the arrival of an inference, we don't model the reasoning process behind it. Agents are already assumed to know all these possible chains of reasoning (concerning not only the knowledge about themselves but also those of their opponents); it is only the game theorist who performs or discovers the underlying reasoning.

Beliefs, on the other hand, are supposed to have a different nature in the sense that beliefs can be contradicted by evidence that is not available to the agent. Notwithstanding, beliefs will be assumed to fulfill consistency requirements in the sense that if something is compatible with our beliefs, it must be possible for this statement to turn up to be true without forcing us to give up any of our beliefs.

Unless otherwise stated, the analysis followed in the present work is the logic of knowledge and belief developed in Hintikka [12].

For the reader who is willing to skip the technical aspects explained in the remainder of section 2.3 there is a summary at the end of the section.

2.3.1 Knowledge and the rules of consistency

We assert that a statement is *defensible* if it is immune to certain kinds of criticisms. Knowing *p and n*ot knowing *q* when *q* logically follows from *p,* will be defined as indefensible. Indefensibility alludes to a failure (past, present or future) to follow the implications of what it is known. This is the notion that will be used from here onward. In other words, if somebody claims that he does not know a logical consequence of something he knows he can be dissuaded by means of internal evidence forcing him to give up that previous claim about his knowledge. Therefore, within the present system of axioms, *logic* has epistemic consequences and this entails that the subjects of the epistemic operators possess logical omniscience. Hintikka doubts that the incapability of having logical omniscience should be defined as *inconsistency*. He proposes the term *indefensibility* to substitute it because, in his opinion, not knowing a logical implication of something we know should not be defined as inconsistency.

In order to define the notion of defensibility we need to introduce the concept of a model set.

Definition: A set of sentences μ is a *model set* iff satisfies the following conditions:

(C.~) If $p \in \mu$, then not $''{\sim}p'' \in \mu$. That is, a model set can not have as members a proposition together with its negation.

(C.&) If $''p\&q'' \in \mu$, then $p \in \mu$ and $q \in \mu$. The elements of a conjunction that belongs to a model set should belong as well.

(C.v) If $''pvq'' \in \mu$, then $p \in \mu$ or $q \in \mu$ (or both). The elements of a disjunction that belongs to a model set should belong as well.

(C.~~) If $''{\sim}{\sim}p'' \in \mu$, then $p \in \mu$. If the double negation of a proposition belongs to a model set, then the proposition should also belong to the model set. To complete the description the De Morgan's rules for negation of conjunction and disjunction need to be introduced:

(C.~&) If $''{\sim}(p\&q)'' \in \mu$, then $''{\sim}p'' \in \mu$ or $''{\sim}q'' \in \mu$ (or both).

(C.~v) If $''{\sim}(pvq)'' \in \mu$, then $''{\sim}p'' \in \mu$ and $''{\sim}q'' \in \mu$.

This set of conditions will be named as the "C-rules".

Definition: A set λ of sentences can be shown to be *indefensible* iff it cannot be embedded in a model set. In other words, for λ to be *defensible* there should exist a possible

state of affairs in which all the members of λ are true and this in turn occurs *iff there is a consistent description of a possible state of affairs that includes all the members of* λ. Our goal is to find a framework to characterize a *defensible* (generally called *consistent*) state of mind in terms knowledge and belief of an agent. For instance, when the notion of a model set is applied to an agent's knowledge, we will see that if an agent 'i' knows that proposition 'p' is true, a defensible state of mind of this agent can not include the contradictory of 'p'. By the same token if 'i' knows that 'p' and 'q' are true then 'i' should also know that 'p' is true and that 'q' is true. The C-rules serve the purpose of defining the consistency of players' states of minds.

2.3.2 Possible or Alternative worlds

We have so far spoken about knowledge and belief and briefly defined the operator $"P_i"$. Assume that we have some description of a state of affairs μ and that for all *i* knows in that state, it is possible that *p*. That is, $"P_i\, p" \in \mu$. The substance of the statement $"P_i\, p"$ can not be given a proper meaning unless there exists a possible state of affairs, call it μ*, in which *p* would be true. However μ* need not be the actual state of affairs μ. A description of such state of affairs μ* will be called *an alternative to* μ with respect to *i*. Therefore, in order to define the defensibility of a set of sentences and give meaning to the notion of alternative worlds, we need to consider a *set* of models. Hintikka calls this set of model sets a *model system*. Within this framework the previous condition regarding the existence of alternative worlds can be formulated as follows:

(C.P*) If $"P_i\, p" \in \mu$ and if μ belongs to a model system Ω, then there is in Ω at least one alternative μ* to μ with respect to *a* such that $p \in \mu^*$.

This condition guarantees that p is possible. In other words, if an agent thinks that for all he knows it is possible that 'p' is true, then there has to be an alternative state of mind consistent with the agent's actual state of mind in which 'p' is true. That is, without incurring in a contradiction, the agent should be able to conceive a hypothetical scenario in which 'p' is true.

Hintikka also imposes the condition that everything *i* knows in some state of affairs μ should be known in its alternative states of affairs:

(C.KK*) If $"K_i\, p" \in \mu$ and if μ* is an alternative to μ with respect to *i* in some model system Ω, then $"K_i\, p" \in \mu^*$.

This means that alternative worlds should be epistemologically compatible with respect to the individual whose knowledge we are denoting. Alternative worlds do not lead the agent to contradict or discard knowledge.

Additionally the following conditions needs to be imposed:

(C.K) If $''K_i\, p''$ $\in$ μ, then p $\in$ μ. This says that knowledge cannot be wrong. In other words, if *i* knows that p then p is true.

(C.~K) If $''\sim K_i\, p''$ $\in$ μ, then $''P_i\, \sim p''$ $\in$ μ. This means that it is indefensible for *i* to utter that "he does not know whether *p*" unless it is really possible for all he knows that *p* fails to be the case.

(C.~P) If $''\sim P_i\, p''$ $\in$ μ, then $''K_i\, \sim p''$ $\in$ μ. When *i* does not consider *p* possible then, *i knows* that *p* is not true.

Definition: a *model system* is a set of sets that satisfies the following conditions:

i) each member behaves according the C-rules, (C.K), (C.~K) and (C.~P).

ii) there exists a binary relation of alternativeness defined over its members that satisfies (C.KK*) and (C.P*).

2.3.3 The relation of alternativeness

I can be shown that (C.KK*) and (C.K) together imply:

(C.K*) If $''K_i\, p''$ $\in$ μ and if μ* is an alternative to μ with respect to *i* in some model system Ω then $p \in \mu^*$.

In other words, if *i* knows that *p* in his actual state of mind, then *p* must be true not only in that world but also in *any* alternative world with respect to *i*.

Under (C.K*), condition (C.K) can be replaced by:

(C.refl) The relation of alternativeness is reflexive.

That is, every world is an alternative to itself. From this it follows that:

(C.min) In every model system each model set has at least one alternative.

Moreover (C.min) together with (C.K*) imply:

(C.k*) If $''K_i\, p''$ $\in$ μ and if μ belongs to a model system Ω, then there is in Ω at least one alternative μ* to μ with respect to *i* such that $p \in \mu^*$.

The condition of transitiveness also holds for this binary relation and it is implied by the other conditions (for the proof see Hintikka [13] page 46).

The alternativeness relation is reflexive, transitive but not symmetric. To see why the symmetry does not hold consider:

$\mu = \{\ ''K_i\, p'', p, ''P_i\, u''\ \}$

$\mu^* = \{\ ''K_i\, p'', p, ''K_i\, h'', h\ \}$

μ* is an alternative to μ with respect to the individual *i* because the state of affairs in μ* is compatible with what *i knows in* μ. Assume that *u* entails ~*h*. The additional knowledge in μ* is not incompatible with the knowledge in μ but with what *i* considers

possible in μ. However given that *h* entails ~*u*, then μ is not an alternative to μ* (see Hintikka [12] page 42).

To conclude, we say that a member of a model system is *accessible* from another member, if and only if we can reach the former from the latter in a finite number of steps, each of which takes us from a model set to one of its alternatives.

The different sets of rules that are equivalent to each other and that completely define the notion of knowledge are as follows:

(C.P*) & (C.~K) & (C.~P) & (C.K)&(C.KK*)

(C.P*) & (C.~K) & (C.~P) & (C.K)&(C.K*) &(C.trans)

(C.P*) & (C.~K) & (C.~P) & (C.refl) & (C.K*) & (C.trans)

(C.P*) & (C.~K) & (C.~P) & (C.refl) & (C.K*) & (C.KK*)

2.3.4 Belief and the rules of consistency

We can replace all the previous conditions with the exception of (C.K) by replacing the operators "K" and "P" for "B" and "C" respectively. The condition (C.K) does not have a doxastic[7] alternative because it expresses that whatever somebody knows has to be true, which by definition obviously does not hold in the case of beliefs. We already stated that (C.refl) is a consequence of (C.K*) and (C.K). Therefore the reflexiveness does not hold in the case of beliefs. The condition that is valid for beliefs and that will be used here is the following (C.b*), which is the counterpart of (C.k*):

(C.b*) If $"B_i\, p"$ ∈ μ and if μ belongs to a model system Ω, then there is in Ω at least one alternative μ* to μ with respect to *i* such that $p \in \mu^*$.

If *i* believes that *p*, then there is a possible world alternative to the actual with respect to *i* in which *p* is true.

The different sets of rules that are equivalent to each other and that completely define the notion of belief are as follows:

(C.b*)&(C.B*)&(C.BB*)

(C.b*)&(C.B*)&(C.trans)

In the remaining sub-sections we characterize the interaction of knowledge and belief. This is necessary because the players' states of minds will combine these two different operators. We will for instance assume that players have knowledge about the rules and structure of the game but we will only assume that they possess beliefs concerning out-of-equilibrium play. The extent to which rationality can be *known* will be addressed in section 3.

---

[7] A doxastic alternative is an alternative in terms of *opinion* not in terms of knowledge.

2.3.5 The interaction of the knowledge and belief operators

The alternatives to which the knowledge operator applies will be called *epistemic alternatives* whereas the ones to which the belief operator applies will be called *doxastic alternatives*. To be more precise, these denominations should correspondingly replace the previous notions of "alternative".

Definition: an *epistemic (doxastic)* alternative to an actual state of affairs is a description of a state of affairs that is *knowledge(belief)-consistent*.

Once this difference between alternatives in terms of knowledge and belief has been acknowledged, it is easy to see that some conditions that hold for epistemic alternatives do not hold for doxastic alternatives. We already saw that (C.refl) failed to hold for the belief operator what means that it does not hold for doxastic alternatives.

In addition, consider the following condition:

(C.KK* dox) If $"K_i\, p"$ $\in$ μ and if μ* is a doxastic alternative to μ with respect to *i* in some model system Ω then $"K_i\, p"$ $\in$ μ*.

In other words, every world which is an alternative in terms of *i*'s opinion should be compatible within *i*'s knowledge.

This condition can be shown to be equivalent to:

(C.KB) If $"K_i\, p"$ then $"B_iK_i\, p"$ $\in$ μ. That is, *whenever one knows something, one believes that one knows it*. Moreover within the present system whenever one knows something one knows that one knows it. That is $"K_iK_i\, q"$ is equivalent to $"K_i\, q"$. Therefore, all the rule (C.KB) establishes is that *whatever one knows one believes it*. In other words, if $"K_i\, q"$ then $"B_i\, q"$ $\in$ μ.

Moreover, (C.KB) also carries the logical omniscience assumption in the sense that whatever follows logically from our knowledge should be believed: it would be *indefensible* not to believe something that logically follows from our knowledge. Therefore, (C.KB) and (C.KK* dox) will be accepted as conditions.

An interesting feature is that the following rule can not be accepted because it would imply that beliefs can not be given up:

(C.BK) If $"B_i\, p"$ $\in$ μ then $"K_iB_i\, p"$ $\in$ μ. This condition is equivalent to (C.BB*epistemic) and requires that whenever one believes something one knows that one believes it. We assume that by gathering more information one can give up beliefs but not knowledge.

2.3.6 Self-sustenance

So far, we have defined the concept of defensibility as a feature of a set of propositions. The notion of self-sustenance alludes to the *validity* of statements.

definition: A statement p is *self-sustaining* iff the set $\{"\sim p"\}$ is indefensible. Therefore, $"p \supset q"$ is *self-sustaining* iff the set $\{p, "\sim q"\}$ is indefensible.

If $"p \supset q"$ is *self-sustaining* we say that p *virtually implies* q. When p virtually implies q and vice versa then p and q are *virtually equivalent*. In this case, note that $"K_i\, p \supset K_i\, q."$ is *self-sustaining* what means that if a knows that p and pursues the consequences of this item of knowledge far enough he will also come to know that q. In addition, it can be proved that under the proposed set of rules $"K_i\, p \;\&\; K_i\, q"$ *virtually implies* $"K_i\, (p \;\&\; q)"$.

Moreover, within this framework it can be proved that $"K_i\, K_i\, p"$ and $"K_i\, p"$ are *virtually equivalent* whereas $"B_i\, p"$ *virtually implies* $"B_iB_i\, p"$ but not vice versa (Hintikka [13] page 124).

2.3.7 Common Knowledge and Belief

The previous knowledge operators can be replaced by higher degrees of knowledge operators without invalidating any of the accepted rules. This is due to the fact that $"K_i\, K_{i'}\, p"$ and $"K_{i'}\, p"$ are *virtually equivalent* for all i and i'. The common knowledge operator will be denoted by "*ck*" and $"ck\, p"$ will be read as: "there is common knowledge that p."

The common knowledge operator can also be defined as the limit of a mutual knowledge operator of level k where k goes to infinity. In the case of two individuals the mutual knowledge operator can be defined as: $MK^{k}_{(i,i')} \equiv (K_i\, K_{i'} ... K_i\, p) \& (K_{i'}\, K_i ... K_{i'}\, p)$ where each parenthesis has 'k' knowledge operators.

Common belief (cb) is equally defined in spirit but it does not result from the mere substitution of the knowledge operator by the belief operator on the previous formula. This is because within this framework to believe that one believes does not imply that one believes it. Therefore common belief should be defined in terms of the conjunction of all the degrees of mutual belief and can not be reduced to an expression like $MK^{k}_{(i,i')}$.

Summary of section 2.3

In section 2.3, we have defined the conditions under which an agent's state of mind is *defensible*. A defensible state of mind for a player '*i*' can be briefly defined as a *set of propositions* that *represent i's knowledge and beliefs* such that '*i*' *does not contradict*

himself. For instance, a player's state of mind is indefensible when he asserts he does not know a logical consequence of some proposition he claims to know (remember that players are supposed to have logical omniscience). Other examples of indefensible states of minds are: i) the ones that include 'p' and '~p' , ii) the ones that contain 'p&q' but do not include either 'p' or 'q' or both, iii) the ones that contain 'p or q' but neither 'p' nor 'q', etc.[8]

As we already stated, the main difference between *knowledge* and *belief* is that only the former can not be contradicted by observation. *What a player claims to know needs to be true*. In addition, it also follows from Hintikka's logic that *when a player knows something then he believes it*. However, the contrapositive is not true: *a player may believe something without knowing that he believes it* (otherwise beliefs could not be given up).

We have also introduced the notion of *alternative worlds* to represent players' conjectures regarding hypothetical scenarios given their actual state of knowledge and belief. The conditions that these alternative worlds need to satisfy are the following: existence: i) if some proposition is considered possible for all an agent knows, then there should exist at least one alternative world compatible with the actual state of mind of this agent where this proposition is true, ii) if an agent believes that a proposition is true, then there is at least one alternative world compatible with the knowledge he possess in his actual state of mind in which the proposition is true. Preservation of knowledge: iii) whatever is known in the actual state of mind should be known in every alternative world.

To conclude, the *common knowledge* operator has been defined as usual. The sets of rules of consistency or defensibility are naturally extended to higher degrees of knowledge given that within the present language formulas can always be extended by the application of additional knowledge operators. Consider for instance the proposition "player i knows that p", which is true in player i's state of mind. Within the present framework, every alternative world with respect to 'i' should be such that this proposition is true in it. The same would occur to the proposition "player i knows that player j knows that p" if this proposition also belonged to i' actual state of mind.

The notion of *mutual belief* has also been introduced in the same spirit as the mutual knowledge operator. That is, mutual belief of degree 'n' is defined as: everybody believes that everybody believes that everybody... and so on, repeating the operator "everybody believes" 'n' times. It is worth noticing that within this framework to believe that one believes something does not imply that one believes it. Nevertheless, if the mutual belief operator is defined as the conjunction of the different degrees of knowledge then we can

---

[8] 'p' and 'q' are formulas within our language L. For instance these are constructions of the following form: "player 1 takes the money at node 1", "player 2 knows that player 1 knows that player 2 would have taken the money had node 2 been reached" etc.

obtain implications of the following form: if everybody believes that everybody believes then, everybody believes.

2.4 The backward induction algorithm

Before stating the definition of the backward induction algorithm we need to introduce the following concepts:

1) Let $G(h^{n'})$ denote the game that given the public history begins at node n'. The payoff functions in this game will be $u_i(P(z_{n+1}))$ for $n \geq n'$ n=1,2,3. $P(z_{n+1})$ is the final story of the game that finishes at the terminal node z_{n+1}. A strategy profile *s* of the whole game induces a strategy profile $s/h^{n'}$ on any $G(h^{n'})$ in the following way: for each player i, $s_i/h^{n'}$ is simply the restriction of s_i to the histories consistent with $h^{n'}$.

2) A Nash equilibrium is a strategy profile that satisfies the following requirement:

$u_i(s_i,s_{-i}) \geq u_i(s'_i,s_{-i})$ for all s'_i.

The centipede game under consideration has two Nash equilibria: (t_1t_3,t_2) and (t_1l_3,t_2).

The backward induction equilibrium can be defined in the following way:

Definition: a strategy profile *s* of a finite extensive form game with perfect information is a backward induction equilibrium if for every $h^{n'}$, the restriction $s/h^{n'}$ to $G(h^{n'})$ is a Nash equilibrium of $G(h^{n'})$ (Fudenberg and Tirole [10]).

The backward induction solution in our game is (t_1t_3,t_2) and the standard argument to support it can be represented as follows:

Under the assumption of common knowledge of subgame rationality we can assert that

$\{"3" \Box\!\!\rightarrow "t_3"\}$ is true.

This implies that,

$\{"2" \,\&\, \{"3" \Box\!\!\rightarrow "t_3"\} \Box\!\!\rightarrow "t_2"\}$ is also true,

and therefore,

$"1" \Rightarrow ("t_1t_3" \,\&\, "t_2")$

Let us denote the two previous counterfactuals by C_3 and C_2 respectively. Under the assumption of common knowledge of rationality the truth of C_3 implies the truth of C_2 and therefore the play of "t_1" by the root player. We start at the last node by solving C_3. In the next step we consider C_2, the counterfactual at the predecessor node. The link between these steps is that C_3 should be part of the set of true propositions or statements that conjoined with "2" determine the truth of C_2. In other words, player 2 would have taken the money at the second node only if he thought that the money would have been taken at the third node. As we can see, the crucial issue we have to address is whether C_3 and C_2 are simultaneously true.

With this purpose, we construct a test for the backward induction equilibrium by considering strategies as contingent events and then introducing a theory to solve these counterfactuals. *Within each of the approaches* considered in the next section, the equilibrium strategies will result as the outcome of the solutions to these subjunctive conditionals. This result will depend upon the payoff structure and the knowledge and beliefs that players commonly hold at all possible nodes.

3. The backward induction solution and the theories of counterfactuals

In this section two different theories of counterfactuals are interpreted to analyze the backward induction algorithm.[9] Before doing so, a few preliminary issues should be addressed.

The *factual* or *actual* world is the world where players play the backward induction solution. In this world, player 1 takes the money at node 1 and the game ends without player 2 being called to play. Nevertheless, players need a plan or a hypothesis in *this* world for the counterfactual scenario in which at least one of the remaining nodes is reached. Two issues need to be solved concerning this matter: i) the epistemological status of this contingent play and ii) the truth condition of the its hypothetical conjecture. Can there be any mutual *knowledge* concerning player 2's strategy? and, is it true that "*had she had the chance to play, she would have played* t_2"? To answer these questions we first need to consider whether players can *know* that their opponents are rational or that they will behave in a certain way.

3.1 The concept of rationality

The task of making compatible the assumption of rationality with the occurrence of deviations, so that these do not in itself imply a contradiction, requires a definition of rationality capable of capturing contingent play. With this aim, we consider the existence of three levels of rationality. Rationality as a *capability of reasoning* will be understood as maximizing behavior subject to exogenous beliefs. This will be defined as rationality *ex ante* to stress the idea that beliefs need not be correct. Rationality *ex post* will enlarge the concept of rationality ex ante by incorporating a *process for belief formation or updating that is rational,* in the sense of being free of contradictions. Finally, the third level of rationality alludes to the capability of acting upon decisions. To be rational in this last sense simply entails the absence of mistakes.

---

[9] For a detailed presentation see Lewis [14] (alternatively Lewis [15] (pages 57-85)) and Bennett [3].

Definition: a player is *rational ex ante* if he plays a best response given his beliefs, or hypotheses about his opponent's play, whatever those beliefs are.

Definition: a player is *rational ex post* if he plays a best response given rationally formed beliefs or hypotheses about his opponent's play. "Rationally formed beliefs" means that the players have the capacity of *correctly* hypothesize about their opponent's contingent play given their own knowledge, a behavioral assumption and a background theory which is commonly held. This means that the set that represents each players' state of mind should be a *defensible set*.

There is another important concept that is necessary to consider in extensive form games. This is the concept of *node rationality*. The aim is to separate rationality at different nodes, because a player who observes a deviation needs to conjecture about the rationality of his opponents at *future* nodes. The information he receives after a deviation *may* have some implications about further node rationality. This will depend on the theory of counterfactuals that the player is using.

Definition: player 'i' is *rational at node* $n \in N_i$ if he plays a best response given the history of previous play h^n and his *hypotheses* or *conjectures* about future play.[10] This is a type of *ex ante* rationality in the sense that a player may deviate and still be node-rational at that node.

Definition: player 'i' is *subgame rational* if he is rational at node n, $\nabla\, n \in N_i$.

Definition: player 'i' is *fully rational* iff he is ex-post subgame rational and does not make mistakes.

3.1.1 Knowledge and Rationality

The relationship between rationality and observation is a difficult matter to establish. We think on the one hand that there can not be *knowledge* concerning actions that are not actually played in equilibrium and therefore, there cannot be mutual knowledge of *full* rationality. In other words, if knowledge of rationality is conferred by observation then there can not be knowledge of rationality at all nodes if some of them are not reached under equilibrium. On the other hand, a player may play in a way his opponent defines as "rational" by pure error and therefore, *observation* would not necessarily provide enough information to establish this type of knowledge. It is clear at this point that either a bayesian

[10] "*hypotheses*" here stands for: how the player evaluates counterfactuals about future nodes based upon the play that has led to his/her node and some *a priori* or primitive assumptions about the rationality of the opponent.

view is adopted, so that every possible explanation of an observation is given a positive probability, or an assumption is introduced so as to narrow down the indeterminacy of this relationship. We will not allow for mistakes as a behavioral assumption within the equilibrium world. Mistakes might only happen in non-equilibrium words.

Moreover, players need knowledge or beliefs *a priori,* regarding the rationality of their opponents and their opponents' conjectures, because this can not be obtained from experience *within* the game that is about to be played. The concept of *ex ante* rationality was introduced to provide a notion weak enough so that *knowledge* may be justified. One could think that players *might* know that their opponents maximize given their beliefs whatever they are. The goal at this respect, is to resemble the typical assumption of common knowledge in order to match our results with those in the literature.[11] *Ex ante* rationality alludes basically to a *capacity* and to have knowledge concerning the *ex ante rationality* of a player, means to know that he is a maximizer, that is, that he has the *capacity* of choosing the action that optimizes his payoff given his beliefs. This assumption can be only justified in very special cases and for this reason we will also deal with the case of common belief in node rationality. Notice that this capacity to *decide* does not mean that the player will actually *perform* what he chooses. This is what we defined as *full* rationality.

In addition, we need to consider that information might be updated as the game evolves and that this might involve a "change in knowledge" within the game, even within the introspective framework we are dealing with in the present work. Clearly when a deviation occurs players acquire a new piece of "unexpected" information. The concept of or *ex post* rationality is relevant at this respect because it involves the complete chain of reasoning. A player that deviates might be *ex ante* rational. However, if there is no consistent set of beliefs that support the deviation he or she will be considered *ex post* irrational.

Regarding the truth condition of the contingent reasoning involved in equilibrium, players may hold beliefs about these conditionals based upon their mutual knowledge or belief of a *primitive behavioral assumption* plus some *theory* of how to infer conclusions regarding the observation of non expected phenomena. The counterfactual occurrence of a deviation will provide in itself an information to which the corresponding theory of counterfactual should attach some value.

We will assume that there is *common* knowledge of the framework or theory that players use to analyze hypothetical scenarios as a necessary condition to justify an

[11] *Belief* in any of these types of rationality can be easily justified in the sense that players may believe that their opponents are rational as long as they do not confront a piece of observation that assures them that this is impossible.

equilibrium outcome. Whichever theory of counterfactuals is used to analyze an equilibrium notion, it *needs to be at least mutually believed or held amongst the players*. This implies that to have a well founded equilibrium concept in games with perfect information, there should be some mutual agreement regarding the principle by which beliefs at *all* nodes are updated (whatever this principle is).

In the following two sections, we present our argument based upon Lewis' and Bennett's approaches to counterfactuals. In the apendix, a formalization of these results within Hintikka's semantical system will be presented.

3.2 Lewis' theory of counterfactuals

Lewis' theory is based on two fundamental concepts: i) the asymmetric openness of time and ii) the notion of possible worlds.[12]

The first notion can be summarized by the idea that the future is counterfactually dependent on the present, whereas the past is counterfactually *independent* of it. Although the past as well as the future are unique under Lewis' assumption of determinism, the past of the factual world provides an information that the future does not contain and that the present should relax so as to produce the occurrence of the counterfactual antecedent.

The second notion is that of the possible world. This is an epistemological entity; a scenario that despite its actual possibility can be conceived within our mind's framework. In terms of the semantical system presented in section 2, a world is defined as a defensible set of sentences that state what the player knows, believes and thinks it is possible (compatible) given his knowledge (beliefs). An alternative world to the actual world with respect to 'i' given a proposition 'p', is a possible world which is knowledge-compatible with the actual world with respect to 'i', and one in which 'p' is true (see section 2). In our case, the actual world is the world where the players play (t_1t_3,t_2); 'p' could be 'l_1', that is, "player 1 left the money at node 1" or 'l_2', that is, "player 2 left the money at node 2". Possible l_1-worlds are worlds where player 1 left the money at node 1. These are worlds where either (l_1t_3,t_2), (l_1t_3,l_2), (l_1l_3,t_2) or (l_1l_3,l_2) are played.[13] In every l_1-world, players have certain knowledge and beliefs whose consistency and closeness to the (t_1t_3,t_2)-world we are going to examine.

Lewis assumes that there exists a primitive relation of comparative similarity amongst possible worlds. Despite the fact that the principle that defines this ordering is constructed upon our experiences and therefore context dependent, Lewis assumes that whatever this

---

[12] For more detailed exposition see [14] and [15].

[13] Possible l_2-worlds are worlds where player 2 left the money at node 2 and where player 1 left the money at node 1 (worlds, where node 3 has been reached); These are worlds where either (l_1t_3,l_2) or (l_1l_3,l_2) are played.

principle is, it is sufficiently well developed to allow communication between people. The impreciseness of the closeness relationship is due to the intrinsic nature of counterfactuals and possible worlds theories are subject to this criticism.

Lewis describes four types of worlds or counterfactual scenarios:

The first world, w_1, is in matters of facts similar to the actual world, w_0, until shortly before the deviation is supposed to occur. At the antecedent time (t_P) "the deterministic laws of w_0 are violated at w_1 in some simple, localized, inconspicuous way. A tiny miracle takes place." ([14] page 44). The occurrence of the deviation does not necessarily imply that the player chose to deviate. It only implies that he *did* it. A "miracle" is a metaphor to conceive this thought experiment. No further "miracles" occur and after t_P. At w_1, a miracle could be represented by a mistake that produces the corresponding miraculous deviation.

The second world, w_2, contains no miracles. The deterministic laws of w_0 hold throughout the whole domain. Given that these two worlds differ at least in the occurrence of the deviation and have the same 'laws', then it must be the case that they do not agree in matters of particular facts neither before nor after the occurrence of the antecedent. In this case, no miracle produces the deviation. In terms of game theory, off the equilibrium play must arise as the consequence of an *intended* action. However, if players are still rational, which is the assumption we want to consider, it ought to be that their beliefs *justified* that deviation. To reconcile "rationality" with "deviations" we introduced the definition of node-rationality and ex-ante rationality. Otherwise, a deviation would in itself be a contradictory or *impossible* event and this would render all the counterfactuals *vacuously* true providing an inappropriate foundation.

The third world, w_3, has perfect match in terms of facts with w_0 until the deviation. At that time a miracle causes the corresponding off the equilibrium play. Immediately after, a small miracle takes place so as to make the consequent of the counterfactual false.

The fourth world, w_4, is alike w_0 until the deviation obtains. After t_P a widespread second miracle occurs that erases the effects of the deviation in such a way that the consequent is false and no traces of the antecedent deviating play are found.

Lewis' theory: $P \mathrel{\square\!\!\rightarrow} Q$ is true iff either (1) there are no possible P-worlds (in which case $P \mathrel{\square\!\!\rightarrow} Q$ is vacuously true) (2) The closest P-world to the actual world, w_0, is a Q-world or (3) when there is no unique closest P-world, some P.Q-world is closer to w_0 than any P.~Q-world.

To apply this criteria we need to define or impose some ordering of the worlds.

Lewis defines the closeness relationship in accordance with his requirement of the asymmetry of counterfactual dependence by offering a ranking of miracles. As it can be seen, there is a trade off between facts and miracles and the closeness or similarity criteria. The longer the region of perfect match the bigger the miracle we need to produce the

antecedent and vice versa. On the other hand, the farther away in the past the factual discrepancy occurs the smallest the required miracle. In the limit a complete divergence of facts until minus infinity can bring a counterfactual world without the need of a miracle. Compare for example w_1 with w_2. No miracles are allowed in w_2 what means that under determinism these worlds have never coincided in the past. The deviation lawfully occurs due to some different belief.

In Lewis' opinion a world like the previous w_1 will be the typical candidate for the closest world because "a lot of perfect match of particular fact is worth a little miracle" ([14] p. 45). In Lewis' theory worlds would be ranked from the closest to the farthest in the following way: w_1,w_2,w_3,w_4. We'll come back to this discussion because Bennett's theory does not allow for miraculous worlds, so that only type-w_2-worlds are considered.

The asymmetry of counterfactual dependence also brings the result that the miracle at w_4, that produces the reconvergence to w_0, is bigger than the one that produced the divergence. Given that the past is fixed, we need a broader miracle to erase every consequence of the divergent miracle. Therefore, w_4, that contains one small divergent miracle and one big reconvergent miracle, ought to be less close to w_0 than w_3 for this last world contains two small miracles. On the other hand, w_1 is ranked closer to w_0 than w_3 because it contains only one small miracle. In matters of facts, w_2 is the farthest from w_0. The complete absence of miracles can only be gained by a total divergence of the past. However, w_2 is ranked farther from w_0 than w_1 due to the assumed independence of counterfactuals with respect to the past.

The asymmetric openness of time together with Lewis' bias towards the importance of facts previous to t_P, allows fixing the facts or parameters that we want to keep constant to analyze the counterfactual hypothesis. Within w_1, the exogenous variables will be the players' intentions concerning their rational play. Therefore deviations will not imply a revision to the belief that players are node rational at future nodes. Within this world *players do not intend to deviate*, the occurrence of a deviation is miraculous in the sense that it constitutes a *thought experiment* that captures all the features of the actual world with the exception of the deviation; it only affects the map from decisions to actions in the hypothetical case of a deviation but *not in the actual world*.

Let us start by examining C_3 under the assumption of common belief of node rationality.[14] We first need to search for the closest hypothetical world where a deviation occurs. Within Lewis' paradigm, the smallest miracle that can produce a "3"-world (a world

[14] Within our interpretation of this theory, common *knowledge* of node rationality yields the same results. Moreover, we could have as well assumed common *belief* in the theory to analyze counterfactual scenarios instead of common knowledge.

where node 3 is reached), without bringing an inconsistency between rationality and deviations, is one in which some tremble caused the previous players to leave the money. No further miracles are allowed so as to *resemble behavior in the actual world as close as possible*. Under any definition of rationality[15] this possible "3"-world is a "t_3"-world given that new miracles are ruled out (this means that the player at the last node can not make a mistake). Furthermore, this world, call it w_R, is under Lewis' metric the closest to the equilibrium world w_0 that allows us to assert the truth of C_3. Note that there could be other worlds different from w_R in which "3" is true. Clearly the case in which players 1 and 2 are both irrational (name it w_I). In this case, player 1 would play l_3 so that the counterfactual C_3 is false.[16] However, under the assumption of rationality, w_R should be closer to the equilibrium world than w_I what brings C_3 true. The world of trembles is another type of world where C_3 *may* be false although with probability approaching zero. Note again that the deviation does not occur in the *actual* world. However, under this framework a possible let us say, l_1-world is a world where no trembles are *further* expected.[17]

Under our interpretation of Lewis' theory, deviations are not incompatible with players' *ex post* rationality because players *need not behave irrationally again after a deviation*. Deviations are incompatible with *full* rationality. Nevertheless, there is a difference between the counterfactuals C_3 and C_2 in terms of the informational structure needed to support them and their connection to rationality. At the end of the game, expectations about the opponent's rationality do not count[18]. But this is not the case at node 2, where the task of making compatible the assumption of rationality with the occurrence of the necessary previous deviations (so that it does not in itself imply either a contradiction or the expectation that C_3 is false) requires a definition of rationality capable of capturing contingent play. With that aim, the concept of node rationality above stated is to be used at this stage. Under this definition, a player may deviate and still be node-rational. This feature will be crucial within the next framework where intentional play is assumed.

It has been already asserted that at the second node, the truth of C_3 can not be *known* to player 2 (remember that we are assuming common belief); however, the truth of C_3 together with the truth condition of any other counterfactual can be *hypothesized* on the

---

[15] Irrationality it is not defined here as a particular case of rationality. Rationality and irrationality are meant to be two disjoint categories. Given some beliefs at a node, the rational behaviour is to choose the action that yields the highest payoff what is a trivial problem at the last node since there are no ties in this game.

[16] $(P \mathrel{\square\!\!\rightarrow} Q)$ is false iff $(P \mathrel{\square\!\!\rightarrow} \sim Q)$ is true. See [15].

[17] Within the trembling hand refinement the probability of trembles goes to zero within the actual or equilibrium world. Outside this world, trembles at every node are possible but uncorrelated.

[18] It is assumed here that players have no uncertainty regarding all the payoffs in the game.

basis of the *common belief of node-rationality*, that is all that will be required in the present analysis. As a consequence of this assumption the truth of C_3 is commonly believed.[19]

Consider now C_2. We have to establish whether "t_2" is true in the closest ["2"&C_3]-world. In this world, a miracle produces the play of l_1 by player 1 so that player 2 gets the chance to play. The decision at that node will depend on what he expects player 1 to play at the third node. To begin with, we have to find a world in which the conjunction ["2"&C_3] is not false. Consider first the following candidates for possible "2"-worlds:

w_R: where a miracle consisting of a mistake causes the previous deviation but contains no further breaches of laws,

$w_{R'}$: a world where player 1 is node rational at node 1 but has the wrong beliefs about player 2's rationality,

$w_{R''}$: where player 1's beliefs are right about the irrationality of player 2 and

w_I : where player 1 is the only irrational player.

Notice that in the last three worlds deviations are intentional that is, they do not contain miracles while the first does. The crucial question is: in which of these worlds would C_3 be true?[20]

For expositional purposes let us represent these worlds in terms of the rationality of the players and the events that hold true in them:

World	Player 1	Player 2	Type of world
w_R	subgame rational & mistakes	subgame rational	l_1t_3,t_2-world
$w_{R'}$	subgame rational with wrong beliefs	subgame rational	l_1t_3,t_2-world
$w_{R''}$	subgame rational with right beliefs	subgame irrational	l_1t_3 ,l_2-world
w_I	subgame irrational	subgame rational	l_1l_3,l_2-world

The first three candidates are worlds at which player 1 is node-rational at all nodes, so in any of them C_3 is true. We rule out w_I as a possible closest world within this approach. Now we have to find the closest deviation-world and see whether "t_2" is true in it.[21]

---

[19] Common belief of node rationality is sufficient for the truth of C_3 ; so is common belief of subgame rationality that is an stronger assumption.

[20] Another important question, that we do not address here is which of these worlds is more sensible as an explanation of what went wrong. We do not postulate that players do believe in mistakes, we only look for the logical implications of that assumption which is consistent with our interpretation of Lewis' criterion.

[21] In case of a tie regarding the closeness of the worlds with respect to w_0 , C_2 is true iff a [w_j &"t_2"] world is closer to w_0 than a [w_j &~"t_2"] world, where j denotes the equally distant worlds.

The world $w_{R''}$ can be eliminated because player 2's irrationality ranks it further from the others in terms of *features that should be preserved*. So, we reduce the set of possible worlds to w_R and $w_{R'}$. Although these two worlds are l_1t_3,t_2-worlds it is interesting to see which one is the closest in order to compare it with Bennett's theory of counterfactuals which will be introduced in the next section. First note that $w_{R'}$ is a w_2 type of world. There are no miracles, given that player 1's play is intentionally guided by some beliefs. However, these beliefs are not compatible with the assumption of common belief of players' subgame rationality. To go from w_0 to $w_{R'}$ we need to change a feature of the actual world, that is the belief of player 1 about player 2's rationality which was supposed to have a parametrical role under our assumption of rationality. Lewis does not allow for this change in crucial parameters. Therefore, we are left with w_R where player 2 is supposed to play t_2 given the assumption of node rationality.[22] In this way we obtain the backward induction solution.

There are some relevant issues at this point. It is claimed that the size of the required miracle that produces a deviation up to the last node of the game increases with the number of nodes in this game and that this may disturb the previous ranking.[23] This is a valid and interesting issue and it is beyond the goal of this paper, which deals with a tree-legged centipede game. Nevertheless, even if we consider that correlated mistakes would produce a smaller departure from the actual world, capable of bringing all the deviations that are needed, this will not alter the truth of the counterfactual at the last node *when no further miracles are expected* and *when the last player is node-rational*. This is due to the fact that *every* "3"- world is a t_3-world under our assumptions of rationality. If the truth of C_3 is commonly believed then the previous argument should unravel by backward induction. The key element in this argument is that beliefs are "revised" in such a way that *common belief of node rationality* is still possible *after a deviation*. Assume player 1 plays l_1 believing that player 2 will therefore believe that he is irrational and very likely to leave the money at node 3. Assume also that, player 1 plans to take the money at the third node. All these propositions could be commonly believed *only if* players had incomplete information about the existence of irrational players in the actual world (see Binmore [6], McKelvey and Palfrey [16] and Reny [17]). It should be said that this is another way to model this game and that other equilibria arise. However, it is not the only way to think about deviations.

---

[22] It could be said as it is implied in Binmore [7] that given player 1's deviation now player 2 may expect the play of l_3 at the last node, justifying in this way the play of l_2. However this is still incompatible with the truth of C_3 under the assumption of common knowledge of node rationality when deviations are conceived as thought experiments o small miracles in terms of Lewis' approach.

[23] This is one of Binmore's remarks. See [6]

Given our definitions, *ex ante* rationality can be mutually known among the players if it is possible for them to know this as a feature or capacity of their opponents. That is, if rationality is considered to be a *disposition*. In this case, we can keep the assumption that *actions* are not necessarily *known* to the players. All players might know is their opponent's capacity to optimize given his beliefs. The previous argument also holds if this alternative view of rationality is accepted and knowledge is postulated instead of belief.

3.3 Bennett's theory of counterfactuals

Under Bennett's theory of counterfactuals, the past can counterfactually depend on the future because no miracles are allowed to keep the closeness in facts to the antecedent time. In this case, if something contrary to fact is observed this implies that some previous conditions must have been different for this predicate to have occurred.

As it was asserted, under the assumption of node rationality a player may deviate and still be node and subgame rational depending on the beliefs he holds at the corresponding nodes about future contingent play.

Under our interpretation of Bennett's theory, beliefs are the endogenous variables that support hypothetical play. In Lewis' approach, players are rational and miraculously off-the-equilibrium nodes are reached. Under our interpretation of Bennett's theory, on the other hand, deviations from a certain equilibrium should be explained by beliefs that make this behavior a rational choice.

Definition: It is said that ($P \ \Box\!\!\rightarrow \ Q$) is true à la Bennett if Q is true at all the antecedent time-*c*losest-causally possible P-worlds. That is, we start at t_P, the moment in which the deviation occurs, then we lawfully unfold the facts in both forwards and backward directions. If Q is true in each of these worlds then the counterfactual is true. In other words, once the deviation occurred, we reason backward by finding the corresponding beliefs that the players ought to have had in order to have played node rationally. With these beliefs, we unfold forwards the sequence of facts to see if, in this world, the counterfactual consequent is true. This treatment makes beliefs endogenous and rationality exogenous because the mean by which the deviation occurred is *derived* as a residual instead of being assumed. Therefore, there is no need to assume a theory of mistakes to justify the occurrence of the counterfactual antecedent.

Let us start at a world in which, without any violations to the assumption of rationality, the second node is reached, for we have already seen why C_3 is true under any theory of counterfactuals and any definition of rationality. Starting at a world where the second node is reached, we have to unfold the consequences in both directions of time and see whether "t_2" obtains. At this node, player 2 has to decide whether to play the equilibrium

action or not. This choice should be guided by the expected play at the third node that would have resulted had that node been reached. Bennett's worlds are w_2-type of worlds; in terms of the previously defined worlds, they are worlds like $w_{R'}$ or $w_{R''}$. Following Lewis, in the previous section we ranked w_R as closer to w_0 than $w_{R'}$ or $w_{R''}$. Bennett's approach does not allow for a world like w_R so this case is ruled out. Moreover, we discard worlds like w_I for being farther from the actual world in which players are rational by assumption.

Bennett's criterion and the assumption of rationality lead to the conclusion that had node 2 been reached, then the play of l_1 by player 1 ought to have been motivated by the belief that player 2 would play l_2 at that node. Within our interpretation of Bennett's worlds, deviations are intentional, that is, players are supposed to have had a reason for deviating. However, both players expect that player 1 would have played t_3 had node 3 been reached based on the common belief of node rationality.

Consider first a world like $w_{R'}$. If, due to the commonality of the belief about the play at node 3 and the node rationality of player 2, it is implied that player 2 would have played t_2 had node 2 been reached, then this implies that *either player 1 is node-rational at node 1 and the commonality in the belief of "*t_2*" can not be held, or that player 1 is node irrational at node 1*. Therefore, if player 1 is supposed to be *ex post* rational at all nodes (as the standard argument goes) then the truth of C_2 can not be commonly held. On the other hand, if we keep the common belief on C_2, we have to rule out common belief in subgame rationality.

Consider a world like $w_{R''}$. Player 1 is not mistaken about his beliefs regarding player 2's play and he is node rational at all nodes; however, player 2 is not node rational. In this world player 2 plays or l_2 so that C_2 is not true. But this type of world should be farther than $w_{R'}$ because in $w_{R'}$ both players are subgame rational. [24]

None of these worlds under consideration are compatible with the common belief in the truth of C_2 and in *ex post* rationality. If C_2 is commonly believed to be true, in the hypothetical occurrence of "2", it would be *known* to the players that either player one made a mistake, what is ruled out by assumption, or that he is not *ex post* rational at that node. That is, *there is no consistent set of beliefs that can support this deviation*. Therefore, both counterfactuals can not hold true under this theory *if common belief of ex post rationality is to be assumed at all nodes*. The truth of C_2 is not consistent with the play of l_1 and the assumption of common belief of subgame *ex post* rationality. Players' states of mind are not defensible according to the definition in section 2.3 and this is commonly believed. The key feature that brings this result is the combination of the assumption that players have

[24] In the presence of only one node for a player node and subgame rational are equivalent concepts.

common belief in subgame rationality and that a deviation *necessary* carries a consistent intention.

3.4 The formalization of the results

This section analyzes the conditions under which the backward induction outcome obtains in terms of levels of mutual knowledge and belief and within the semantics presented in section 2. The following definitions establish the concept of node rationality within our language. The axioms, on the other hand, state the structure of the game and players' rules of inference.

Definitions and axioms:

(A1) Structure of the game: payoffs, available actions and order in which players move.[25]

Subgame rationality: player i is subgame rational (R_i) iff he is node rational at *every node* at which he might have the chance to play (R_{in} n$\in N_i$).

(A2) $R_1 \equiv R_{13}$ & R_{11}

(A3) $R_2 \equiv R_{22}$

Node rationality will be defined in terms of contingent play as follows:

(A4) $R_{22} \equiv$ "l_1" $\Box\!\!\rightarrow [(B_2\, t_3 \,\&\, t_2)\ v\ (B_2\, l_3 \,\&\, l_2)]$

In other words, player 2 is node rational at node 2 iff it is true that had node 2 been reached, he would have either taken the money -if he believed that player 1 would take it in the next round if given the chance- or left it otherwise.

(A5) $R_{11} \equiv$ "r"$\Rightarrow (B_1\, t_2 \,\&\, t_1)\ v\ (B_1\, l_2 \,\&\, l_1)$

Player 1 is node rational at node 1 iff it is true that he takes the money when he believes that player 2 would have taken it in the next round or leaves it otherwise.

(A6) $R_{13} \equiv$ "l_2" $\Box\!\!\rightarrow$ "t_3"

Player 1 is node rational at the third node iff C_3 is true. Remember that:

(A7) "l_2" $\Box\!\!\rightarrow$"t_3" $\equiv$ "C_3"

(A8) "l_1" $\Box\!\!\rightarrow$"t_2" $\equiv$ "C_2"

Rules of inference:

(A9) Conditions (C.P*),(C.~K),(C.~P),(C.K),(C.KK*dox)&(C.KB)

Knowledge of the game and rules:

(A10) Common knowledge of definitions and axioms (1)-(9)

Observe that according to (A6) if R_{13} is true then (A7) is true. That is, given the definition of rationality under consideration we assume that C_3 is true under any theory of

[25] This was presented in section 2.

counterfactuals under the assumption that the first player is rational at that node. We define R_{13} in this way because expectations do not matter at the last node and we do not assume mistakes to resemble the highest similarity with the actual world. However, note that, from this set of axioms, we can not assert the truth of the second counterfactual. In order to do this we need to introduce a theory to analyze it.

Notice also that (A4) and (A5) define strategies as contingent constructions.

We need to introduce another axiom stating how deviations might be interpreted:

(A11) Bennett's theory: Under this theory, the play of l_1 would provide a new piece of information to player 2 and would make him consider as possible an alternative world where one of the following three alternatives need to be *included:*[26]

$h^{j} = \{"K_2\, l_1"\,, "B_2 \sim R_{11}"\}$; $h^{j} \subset \mu_2{}^{j}$ (Player 2 believes that player 1 is node irrational at node 1 only)

$h^{jj} = \{"K_2\, l_1"\,, "B_2\, (R_{11} \& B_1\, B_2 l_3 \,\&\, B_1\, R_2)"\}$; $h^{jj} \subset \mu_2{}^{jj}$ (Player 2 believes that player 1 is rational, that player 1 believes that 2 is rational and that player 1 believes that player 2 believes that player 1 is irrational at the third node)[27]

$h^{jjj} = \{"K_2\, l_1"\,, "B_2\, (R_{11} \& B_1\, B_2 t_3 \,\&\, B_1\, l_2)"\}$; $h^{jjj} \subset \mu_2{}^{jjj}$ (Player 2 believes that player 1 is rational and that player 1 believes that player 2 is irrational).

The theory can be expressed in the following way:

Definition: "$K_i(A \,\Box\!\!\rightarrow B)$" iff in the closest alternative state of affairs to player i's actual state of affairs such that K_iA is true, K_iB is also true. Note that by (C.K) if "$K_i(A \,\Box\!\!\rightarrow B)$"$\in \mu_i$ then $(A \,\Box\!\!\rightarrow B) \in \mu_i$. Notice that A and B need only be *possible* sentences, not necessarily true *within the player's actual state of mind*; they only need to be true in the *closest* alternative world. In his actual world, player i only needs knowledge of the counterfactual connection between A and B. What is necessary is that there exists a possible world in which A and B can be known to be simultaneously true. The previous definition could have been stated with the operator B_i replacing K_i. In this case player i would believe that the counterfactual connection is true instead of knowing it.

Trivially at the last node beliefs do not matter and hence any definition of rationality suffices to attach the truth of the corresponding counterfactual. An alternative or possible world where the last node is reached would need to include the following state of affairs:

$\mu^{jv} = \{"K_1\, l_2"\,, "B_1\, [(R_{22} \& B_2 \sim R_{13})\, v \sim R_2\,]"\}$ However player 1 would play t_3 in every possible world in which he is node rational at this last node. These worlds are considered closer to the actual where he is rational. Hence any of our criteria to determine the truth of

---

[26] This means that the following sets are not a full description of the alternative worlds, just a subset of it.

[27] Further levels of knowledge could have been assumed without loss of generality. These are the minimum conditions that explain a deviation.

counterfactuals would render this counterfactual true. A different matter is whether player 2 knows or believes this and whether there could be common knowledge that C_3 is true. As we already said, this will only be the case if he does not expect deviations by the first player to be correlated and attaches to this event either an infinitesimal probability or thinks of this scenarios as thought experiments in the spirit of Lewis' approach.

3.5 Primitive epistemic and doxastic structures

The idea is to start with a primitive information structure (E_1 and E_2) and then complete the set that is defensible for each player given the axioms and their knowledge of it. The final or complete defensible set for each player that will reflect his corresponding state of mind will be denoted by λ_i (i=1,2). The aim is to see whether there exists a defensible set that represents the player's states of mind that is compatible with the truth of the corresponding counterfactuals so that the backward induction algorithm is free of inconsistencies. In other words, we need to test whether there exists μ_i such that $\lambda_i \subset \mu_i$ (i=1,2) where μ_1 and μ_2 are model sets as defined in section 2. As it was explained, players' decisions have already been "taken" within their given states of minds. It is the game theorist who searches in the players' minds for consistency or defensibility.

<u>Case 1</u>:

Assume that there is common knowledge of subgame rationality.

$E_1 = \{"K_1\ (A10)," K_1\ [ck(R_1 \& R_2)]", "K_1\ ck(A11)"\}$; $E_1 \subset \lambda_1$

$E_2 = \{"K_2\ (A10)," K_2\ [ck(R_1 \& R_2)]", "K_2\ ck(A11)"\}$; $E_2 \subset \lambda_2$

Players are assumed to have common knowledge of the structure of the game the rules for belief revision and the logical framework. Moreover they are supposed to have common knowledge of ex-ante node rationality. Conditions (A4) and (A5) fully describe the options opened to rational players and this common knowledge.

First notice that, given the assumption of common knowledge, both players should share the same information structure, that is, $\lambda_1 = \lambda_2 = \lambda$. Therefore from now on we use λ indistinctively. By the same argument we only need to consider one model set μ, such that $\lambda \subset \mu$ can be proved to be defensible.

According to E_1 and E_2, "$[K_i\ (A11)\ \& K_i\ [ck(R_1 \& R_2)]] \supset K_i\ [ck\ (t_3)]$" for i=1,2 is <u>*self-sustaining*</u> with respect to the model set μ. Therefore "$K_i\ [ck\ (t_3)]" \in \lambda$, i=1,2 by condition (C.K). In other words given common knowledge of the definition of rationality, the

assumption that players are rational the counterfactual at the third node becomes true and it is common knowledge that had node three been reached player one would have played t_3.[28]

What about the counterfactual at the second node? Here we need to introduce the assumption that there is common knowledge of axiom (A11).

The fact that "K_i [ck (t_3)]" $\in \lambda$ for i=1,2 entails by (C.KB) that "K_i [ck (B_2t_3)]" $\in \lambda$ for i=1,2 that is that "K_i [ck (B_2t_3)]" is self-sustaining for i=1,2.

Following the reasoning and given the knowledge assumptions, we obtain that "K_i [ck$(B_2\ t_3\ \&\ t_2\ )$]" $\in \lambda$ for i=1,2 and by (C.K) and (C.&) that "t_2" is true. That is the second counterfactual *should be true* for λ to be defensible (that is to guarantee that $\lambda \in \mu$) for i=1,2.

The crucial question is whether this counterfactual is true and whether its truth maintains the defensibility of λ. In other words, the hypothetical world entailed by the counterfactual should be a defensible state of mind that considered an alternative to the actual world.

Now we explore the alternative counterfactual worlds. First we need to consider the alternative worlds that are accessible from μ with respect to each player. Recall that an epistemically (doxastically) alternative world need to be knowledge (belief) compatible with the actual state of affairs that we denoted by μ.

Consider h^j= {"$K_2\ l_1$" ,"$B_2\ \sim R_{11}$"} Given that it is common knowledge that h^j is included in a possible state of affairs μ^j, "$ck[P_i\ (K_2\ l_1 \& B_2 \sim R_{11})]$" $\in \mu$; i=1,2.

The question is $(K_2\ l_1 \& B_2\ \sim R_{11})$ self sustaining? Assume the answer is affirmative.

By (C.P*) there exists μ^j such that "$K_2\ l_1 \& B_2 \sim R_{11}$" $\in \mu^j$ where μ^j is an alternative to μ and μ^j is accessible from μ with respect to both players.

By (C.&) "$B_2 \sim R_{11}$" $\in \mu^j$

However "$K_2 R_{11}$" $\in \mu$ and by (C.KK*dox) "$K_2 R_{11}$" $\in \mu^j$ and therefore by (C.KB) "$B_2 R_{11}$" $\in \mu^j$ which is a contradiction. Therefore we rule out any alternative that contains h^i as an alternative world where the theory can be sustained and the second counterfactual be true.(Note that "$K_1(B_2 \sim R_{11})$" does not contradict player 1's knowledge if he does not know what player 2 knows about player 1.

Consider now h^{ij}= {"$K_2\ l_1$" ,"$B_2\ (R_{11} \&\ B_1\ B_2 l_3\ \&\ B_1\ R_2\ )$"}; $h^{ij} \subset \mu^{ij}$

Assume that " $ck[P_i\ (K_2\ l_1 \& B_2\ (R_{11} \&\ B_1\ B_2 l_3\ \&\ B_1\ R_2\ ))]$" $\in \mu$; i=1,2.

By (C.P*) there exists μ^{ij} such that "$K_2\ l_1\ \& B_2\ (R_{11} \&\ B_1\ B_2 l_3\ \&\ B_1\ R_2\ )$" $\in \mu^{ij}$ where μ^{ij} is an alternative to μ and μ^{ij} is accessible from μ with respect to both players.

By (C.&) "$B_2\ (B_1\ B_2 l_3)$" $\in \mu^{ij}$

However we already saw that "K_i [ck (t_3)]" $\in \lambda \subset \mu$; for i=1,2

---

[28]Recall that strategies are defined as contingent structures.

In particular "$K_2\, t_3$" $\in \mu$ and "$K_i K_2\, t_3$" $\in \mu$ i=1,2

By (C.KB) "$B_2\,(t_3)$" $\in \mu$ and given that players share the state of mind μ, this implies that "$K_i B_2\, t_3$" $\in \mu$; i=1,2.

By the assumption of ck and (C.KB) "$B_i B_2\, t_3$" $\in \mu$ and by the same argument, "$K_2 B_1 B_2\, t_3$" $\in \mu$

By the assumption of ck and (C.KB) "$B_2 B_i B_2\, t_3$" $\in \mu$; i=1,2.

By (C.K*) "$B_2 B_i B_2\, t_3$" $\in \mu^{ij}$ what implies that μ^{ij} can not be an alternative to μ.

In other words, in this world player 1 believes that player 2 does not know that the third counterfactual is true and under the assumption of common knowledge of rationality this is a contradiction.

Consider now $h^{iij} = \{"K_2\, l_1"\,, "B_2\,(R_{11} \& B_1\, B_2 R_{13}\, \&\, B_1\, l_2)"\}$ $h^{iij} \subset \mu^{iij}$

Assume $"ck[P_i\,(K_2\, l_1 \& B_2\,(R_{11} \& B_1\, B_2 R_{13}\, \&\, B_1\, l_2)]" \in \mu$ i=1,2.

By (C.P*) there exists μ^{iij} such that $"K_2\, l_1\, \& B_2(R_{11} \& B_1\, B_2 R_{13}\, \&\, B_1\, l_2)" \in \mu^{iij}$ where μ^{iij} is an alternative to μ and μ^{iij} is accessible from μ with respect to both players.

By (C.&) "$B_2\, B_1\, l_2$" $\in \mu^{iij}$

By (C.b*) there exists μ^{iv} belonging to the same model system Ω such that μ^{iv} is an alternative to μ^{iij} where "$B_1\, l_2$" $\in \mu^{iv}$. Applying again (c.b*) we obtain that there should exist another alternative to μ^{iij}, μ^{v} , such that "l_2" $\in \mu^{v}$.

We assume that "$K_1\, R_2$"$\in \mu$ and therefore by (C.KK*) "$K_1\, R_2$" should belong to any alternative of μ. Therefore "$K_1\, R_2$"$\in \mu^{v}$.

In the state of affairs μ^{v} player 2 would have played l_2 and its rationality which is commonly known would only be compatible with "$B_2\, l_3$"$\in \mu^{v}$ what contradicts (C.~) because "K_2 [ck (t_3)]" $\in \mu^{v}$. That is, there is no alternative state of affairs such that the theory of counterfactuals could be valid and consistent with the players' knowledge. In other words, there is no complete set of sentences where axioms (A1)-(A11) can hold simultaneously so that mentioned set can be embedded in a model set. This result is similar in *spirit* to that obtained for the belief operator in section 3.2.

It seems that what makes this theory self-defeating is the combination of common knowledge of node rationality with full intentionality. Following Bicchieri [5], we reduce the degree of mutual knowledge to find the amount of knowledge that is necessary to guarantee backward induction.

Case 2:

$E_1 = \{"K_1\,(A10), "K_1(R_1)", "K_1\, R_2", "K_1 K_2\, R_1", "K_1\, ck(A11)"\}$; $E_1 \subset \lambda_1$

$E_2 = \{"K_2\,(A10), "K_2(R_2)", "K_2\, R_1", "K_2\, ck(A11)"\}$; $E_2 \subset \lambda_2$

In this case there is common knowledge about the rules and structure of the game but not of node rationality of the players. Player 1 knows that 2 is rational and that player 2

knows that he is rational. It is sensible that this degree of knowledge should suffice to bring the backward induction play. Player 1 has the minimum amount of knowledge that would induce him to take the money at the root. On the other hand player 2 knows that player 1 is node rational at all nodes. This makes him expect the third counterfactual to be true and therefore hypothesize that he would take the money as well.

In case 1 there was no defensible alternative world in which both counterfactuals could be true and where knowledge of node rationality persists after a deviation. In this case we can construct alternative worlds for each player that are epistemically and doxastically defensible:

Player 2 's alternative world: the above type h_2^{iij}-world. Player 2 can consistently believe that player 1 believes that he is not node rational. That is there exists μ_2^j such that $h_2^{iij} \subset \mu_2^j$ where μ_2^j is an alternative world to μ_2 with respect to player 2 and $\mu_2^j \subset \lambda_2$.

Player 1 's alternative world: consider $h_1^{jv} = \{"K_1 l_1", "B_1 (R_{22} \& \sim K_2 K_1 R_{22})"\}$; $h_1^{jv} \subset \mu_1^j$. Player 1 can consistently believe that player 2 is node rational and that player 2 does not know he knows that player 2 is node rational. This means that is there exists $h_1^{jv} \subset \mu_1^j$ where μ_1^j is an alternative world to μ_1 with respect to player 1 and $\mu_1^j \subset \lambda_1$.

Case 3:

$E_1 = \{"K_1 (A10), "K_1(R_1)", "K_1 R_2", "K_1 K_2 R_1", "K_1 ck(A11)"\}$; $E_1 \subset \lambda_1$

$E_2 = \{"K_2 (A10), "K_2(R_2)", "K_2 R_1", "K_2 K_1 R_2", "K_2 ck(A11)"\}$; $E_2 \subset \lambda_2$

The only difference with respect to the previous case is that player 2 has one extra degree of knowledge: he knows that player 1 knows that he is rational. The hypothetical scenario in which a deviation occurs would render his theory of the game inconsistent. This happens because in the hypothetical scenario of a deviation, player 2 cannot give up his knowledge concerning the node rationality of player 1 (this rules out hypothetical worlds that contain type-h^j state of affairs), his knowledge that player 1 should expect him to believe that the third counterfactual is true (due to $"K_2 ck(A11)"$) and finally his knowledge that player 1 knows that player 2 is rational (this rules out hypothetical worlds that contain type-h^{iij} state of affairs). This are the possible type of sources of deviations and none of them can be consistently accepted by player 2 given his knowledge as we saw in case 1.

However, player 1 does know this and therefore expects the second counterfactual to be true due to his knowledge of player 2's rationality and player 2's knowledge about 1's rationality. By assumption in this case player 1 does not know whether player 2 knows that he knows that player 2 is rational. Hence, there is a defensible set representing a world alternative to the actual with respect to player 1 where the contingent play required for backward induction does not lead to any inconsistency.

Case 4:

E_1 = {"K_1 (A10),"$K_1(R_1)$","K_1 R_2","K_1K_2 R_1","$K_1K_2K_1$ R_2","K_1 ck(A11)"}; $E_1 \subset \lambda_1$

E_2 = {"K_2 (A10),"$K_2(R_2)$","K_2 R_1","K_2K_1 R_2","K_2 ck(A11)"}; $E_2 \subset \lambda_2$

It is obvious following the previous reasoning that in this case player 1 knows that 2 faces an inconsistency and that therefore is left with no criterion to play. Backward induction can not be supported in this case and for any higher degree of knowledge.

Before leave the assumption of knowledge of rationality, it is worth noticing that had we modeled deviations according to our interpretation of Lewis' theory, we would have dropped (A11) such that the deviation need not be intentional.[29] Instead we would introduce a new axiom stating the alternative ranking of scenarios under this theory. In Lewis' theory, worlds in which some departure from perfect performance explains the deviation are the closest ones. In this case players' knowledge and beliefs regarding further contingent play need not be revised and therefore the sources of inconsistency founded in cases 1,3 and 4 do not arise.

Case 5:

Now we will assume that there is common belief of subgame rationality.

E_1 = {"K_1 (A10),"$K_1(R_1)$","K_1 [$cb(R_{11})$]","K_1 [$cb(R_{13})$]","K_1 [$cb(R_2)$]","ck(A11)"};

$E_1 \subset \lambda_1$

E_2 = {"K_2 (A10),"$K_2(R_2)$","K_2 [$cb(R_{11})$]","K_2 [$cb(R_{13})$]","K_2 [$cb(R_2)$]","ck(A11)"};

$E_2 \subset \lambda_2$

According to E_1 and E_2, "[K_i (A10) &B_i [$cb(R_1\&R_2)$]] $\supset$ B_i [cb (t_3)]" for i=1,2 is *self-sustaining* with respect to the model set μ_i for i=1,2 that respectively represents players' actual states of mind. We can not assert as before that the third counterfactual is known to be true only that it is commonly *believed* within the players' *actual* state of mind.

For backward induction to obtain, the truth of the second counterfactual should be commonly believed. As before, we consider the alternative counterfactual worlds.

First we need to consider the alternative worlds that are accessible from μ with respect to each player. Recall that an epistemically (doxastically) alternative world need to be knowledge (belief) compatible with the actual state of affairs that we denoted by μ. Although there is common knowledge of the rule for belief updating players need not share the same state of mind in terms of beliefs. Their states of mind should be compatible in terms of knowledge given that knowledge can not be wrong.

---

[29] Recall that under our interpretation of Lewis' theory the hypothesis of a deviation does not lead to the deletion of any feature that can have causal connection with the occurrence of the counterfactual consequent.

Consider h^j={"$K_2\, l_1$" ,"$B_2 \sim R_{11}$"}. Due to *"ck(A11)"*, it is common knowledge that h^j is included in a possible state of affairs $\mu_i{}^j$ and that *"ck*[P_i ($K_2\, l_1$&$B_2 \sim R_{11}$)]" $\in \mu_i$ i=1,2. The question is whether ($K_2\, l_1$&$B_2 \sim R_{11}$) is self sustaining.

Assume the answer is affirmative.

By (C.P*) there exists $\mu_i{}^j$ such that "$K_2\, l_1$&$B_2 \sim R_{11}$" $\in \mu_i{}^j$ where $\mu_1{}^j$ and $\mu_2{}^j$ are alternatives to μ_i and they are accessible from μ_1 and μ_2 respectively due to the common knowledge assumption regarding the update of beliefs in counterfactual scenarios.

By (C.&) "$B_2 \sim R_{11}$" $\in \mu_i{}^j$; i=1,2.

However "$K_1 R_{11}$" $\in \mu_1$ and by (C.KK*dox) "$K_1 R_{11}$" $\in \mu_1{}^j$. By cb(R_{11}), "$B_1 B_2 R_{11}$" $\in \mu_1{}^j$ yet "$B_2 \sim R_{11}$" $\in \mu_1{}^j$. This means that player 1's belief about player 2's beliefs was wrong. What about "$B_1 B_2 \sim R_{13}$"? Player 1 has no reasons to drop this belief. Therefore he should play t_1.

On the other hand "$B_2 \sim R_{11}$" $\in \mu_2{}^j$ and this leads player 2 to abandon the belief that player 1 is rational at that node. However he still believes that player 1 is rational at the third node. So he plays t_2. Nonetheless, the assumption of common belief must be dropped.

Consider now h^{ij}= {"$K_2\, l_1$" ,"B_2 (R_{11}& $B_1\, B_2 l_3$ & $B_1\, R_2$)"}. Due to *"ck(A11)"*, it is common knowledge that h^{ij} is included in a possible state of affairs $\mu_i{}^{ij}$ and that *"ck*[P_i (K_2 l_1&B_2 (R_{11}& $B_1\, B_2 l_3$ & $B_1\, R_2$))]" $\in \mu_i$ i=1,2.

By (C.P*) $\mu_i{}^{ij}$ such that "$K_2\, l_1$ &B_2 (R_{11}& $B_1 B_2 l_3$ & $B_1\, R_2$)" $\in \mu_i{}^{ij}$ where $\mu_1{}^{ij}$ and $\mu_2{}^{ij}$ are alternatives to μ_i that are accessible from μ_1 and μ_2.

By (C.&) "B_2 ($B_1\, B_2 l_3$)" $\in \mu_i{}^{ij}$ i=1,2.

Both players need to drop a belief to reach a world where there is some degree of mutual belief that player 2 believes that the money will be left at the end. In this world player 1 believes that player 2 beliefs that the third counterfactual is not true and under the assumption of common knowledge of any theory of counterfactuals the truth of "t_3" should be known. Therefore no defensible state of mind can be reached in this case by any player.

Consider now h^{iii}= {"$K_2\, l_1$" ,"B_2 (R_{11}& $B_1\, B_2 R_{13}$ & $B_1\, l_2$)"}. Due to *"ck(A11)"*, it is common knowledge that h^{iij} is included in a possible state of affairs $\mu_i{}^{iij}$ and that *"ck*[P_i (K_2 l_1&B_2 (R_{11}& $B_1\, B_2 R_{13}$ & $B_1\, l_2$)]" $\in \mu_i$ i=1,2.

By (C.P*) there exists $\mu_i{}^{iii}$ such that "$K_2\, l_1$ &B_2(R_{11}& $B_1\, B_2 R_{13}$ & $B_1\, l_2$)" $\in \mu_i{}^{iii}$ where $\mu_1{}^{iii}$ and $\mu_1{}^{iii}$ are alternatives to μ_1 and μ_1 that are accessible from μ_1 and μ_2.

By (C.&) "$B_2\, B_1\, l_2$" $\in \mu_i{}^{iii}$

However from *cb(R_{22}),* (c.b*) and the transitivity property there exists μ^{vj} belonging to the same model system Ω such that μ^{vj} is an alternative to $\mu_2{}^{iii}$ where "$B_2 B_1\, R_{22}$" $\in \mu^{vj}$.

In the state of affairs μ^{vj} player 2 would have played l_2 and his rationality which is commonly believed would only be compatible with "$B_2\, l_3$"$\in \mu^{vj}$ what contradicts (C.~) because "K_2[ck (t_3)]" $\in \mu^{vj}$. This means that player 2 cannot access a world in which he is

not supposed to be node rational. Therefore it must be that "$B_2\ B_1\ t_2$" $\in \mu_i$ [vj]. The question is whether μ_i [iii] could be reached from μ_2. Player 2 *should give up some beliefs* in order to have access to that world. As before, players may still play the backward induction outcome but the assumption of common belief of node rationality can not hold in alternative or hypothetical scenarios.

There is no *alternative* state of affairs such that the theory of counterfactuals could be valid and consistent with players' knowledge and the assumption of common belief in node rationality *ex post*. However, there are possible states of minds one for each player reachable from their actual states of minds *where their decision is the backward induction outcome and where common belief in node rationality can be assumed*[30]. Here we need to impose a criteria of closeness to drop hypothesis like h[ii] above. The other option is to relax ck(A11) for cb(A11). If the theory can be relaxed then the inconsistency need not obtain. However this is not a good solution unless we allow for the coexistence of different theories such that when one is dropped an alternative is chosen. The purpose of the present analysis is to compare the performance of these two theories and not to offer a general framework.

3.6 Review of the results

The typical backward induction argument is free of contradictions in the following circumstances:

i) After a deviation, players update their assumptions regarding rationality, according to our interpretation of Lewis' theory of counterfactuals, and we assume *common knowledge* (or belief) of rationality.

Within Lewis' theory, there are worlds in which players deviate being still rational *ex ante* and *ex post*. There are also worlds in which this is not the case; that is, where there is node irrationality. However, under the assumption of common knowledge of *ex ante* rationality, the former type of worlds are closer to the actual than the latter. In the counterfactual world of a deviation, namely, at the node where the deviation occurred, the *a priori* belief that the player was not going to deviate need to be given up. Nevertheless, under the proposed criteria for closeness, this does not lead to reject the belief or knowledge in further node rationality or node rationality at other deviation-worlds. The reason lies in the way in which beliefs are revised. More precisely, deviations need not have *causal* consequences. There are no intentions underlying the deviation, neither a behavioral assumption correlating decisions at different nodes. Deviations may be related to the

[30] A smaller degree of mutual belief is necessary and sufficient. In the present game. Player 1 needs to believe that player 2 is rational and that player 2 believes that player 1 is rational.

performance of the action (which also a form of irrationality) and not with the decision itself. Therefore, rationality along the decision process is kept. Notice that the wrong performance does not take place in the actual but in the counterfactual world.

ii) Players update their beliefs about the intentionality of the deviator (no mistakes are allowed out-of-the equilibrium path and therefore all behavior is intentional) but their *knowledge is limited*. For instance, if player 2 does not know that 1 knows that he is rational, then observing a deviation does not contradict his former knowledge and belief. This can be obtained under the present interpretation of Bennett's theory of counterfactuals. In the version of the centipede game given in this paper, this obtains when player 2 knows that player 1 is rational, player 1 knows that player 2 is rational and player 1 knows that player 2 knows that he is rational. The drawback is to justify why the level of knowledge is exactly the one required. This means that, if the players were to face the game again with one node less (being this possible[31]) the theory would become inconsistent or self-defeating. Therefore, it does not seem to be a robust result.

iii) Players update their beliefs as in ii) but instead of knowledge they have mutual belief in their node rationality. The degree of mutual belief that is necessary has a lower bound for the root player equal to the number of nodes minus one. The degree of mutual belief is equal to the degree of mutual knowledge needed in ii) above (see Bicchieri [4] and Samet [18]).

On the other hand, our analysis leads to a contradiction within the set that represents players' knowledge and beliefs in the following cases:

iv) There is only intentional play (deviations confer information about the intentions of the deviator) and players have knowledge that exceeds the level of knowledge that is necessary for ii) above. In a three-legged centipede game this obtains for any level of information of player 1 in which *at least* he knows that 2 knows that he knows that 2 is rational. This is the lower bound. The upper bound is infinity, which is the case of common knowledge. When player 2 knows that 1 knows that she is rational she knows that there is something wrong with her theory. Therefore she is left with a contradiction. If player 1 does not know this, then backward induction obtains. However, if player 1 knows that this inconsistency results, he also knows that 2 is facing a contradiction and therefore player 1 himself is left with no theory and backward induction fails. For higher levels of knowledge this naturally keeps on holding (see Bicchieri [4]&[5]).

v) There is only intentional play and players have mutual *belief* in node rationality with a degree that exceeds the lower bound defined in ii). The upper bound is again, infinity.

---

[31] Assume we started at least with four nodes.

4 Concluding remarks

1) According to our interpretation of Lewis' theory, deviations do not possess any meaning in themselves; they can be interpreted as *once and for all trembles*. The main reason why this approach provides a proper foundation for the backward induction result is that mistakes or trembles are not correlated. It is worth remembering that Lewis imposes this condition to guarantee the closest resemblance to the factual world. In the original backward induction argument, deviations do not have meaningful consequences, neither in the world in which they occur, nor in other counterfactual scenarios. Within our interpretation of Lewis' miraculous worlds, deviations have a meaning in the world of the deviation under analysis (in this world the player played irrationally) but carry no consequences in terms of behavior at other nodes off-the-equilibrium path. The way in which the hypothetical scenario of a deviation is brought about, is irrelevant to the assumption about further rationality.

2) Under our interpretation of Bennett's framework on the other hand, deviations give some information about the beliefs of the deviator, who might have had a reason to have deviated, given that he always maximizes. In this case, counterfactual worlds are such that *if* C_2 is true, then *either* player 1 is node rational and both players do not commonly believe in the truth of C_2, *or* C_2 is true and commonly held but it is also commonly held that player 1 is *ex post* irrational at the root. This last result in itself does not imply the falsity of C_3 if we differentiate between *rationality in choosing an action* at different nodes and *rationality in belief*. Note that playing l_1 is not node irrational *per se*; player 1 is node irrational *if* he plays l_1 *while* believing that C_2 is true. On the other hand, to leave the money at the last node is *fully irrational* because no beliefs matter and is the play that leads with certainty to the worst payoff at the node.

3) The theories of counterfactuals above presented, reveal that *there is no unique way* to solve the context dependence in which counterfactuals are generally stated. Moreover, the criteria presented in this paper do not exhaust the list of possible ways to model counterfactual scenarios. There are frameworks in which deviations provide some meaningful information to the players: namely, that the deviator may be intrinsically irrational and may play irrationally at other nodes. Binmore [6], McKelvey R. and Palfrey [16] and Reny [17] offer versions of this approach.

Games of incomplete information provide a framework to incorporate irrational behavior that responds to a pattern. In some contexts, for instance a long centipede, they

may offer a more realistic explanation[32]. Consider a game of incomplete information where there are two types of players: maximizers and altruists. Maximizers are bayesian decision-makers and have the payoffs we have so far worked with. Altruists have payoffs that make them choose "leave" at every node where they may be called to play. Players' prior probability of altruists is $p>0$ and it is common knowledge. It is worth noticing that players' prior probability of altruists or irrational types, not only affects counterfactual worlds but also the actual world. For this reason, and depending on p, rational types may under equilibrium leave the money or play mixed strategies at all nodes but the last one, where they always take it. A rational type, who plays at the previous to last node, may mix or play leave, depending on his beliefs regarding the existence of an irrational type at the last node.

A question may be posed at this stage. Shouldn't there be a way to decide which of the theories of counterfactuals is more suitable? The answer to this matter crucially depends on how we think about rationality and on the context and structure of the game. If rationality is considered a human *capacity*, we have to admit that players may make rational choices but for some reason fail to *perform* them. In this scenario, a miracle is a metaphor for thinking about the occurrence of an unintended, uncorrelated deviation. This approach is free of inconsistencies and always yields the backward induction outcome (with its pros and cons).

Within the framework offered by games of incomplete information, like the one described, irrationality is correlated. Therefore, the possibility of irrationality is always open, *also in the equilibrium world*. This feature is embedded in the model through p and the nature or payoffs of the altruists. In games of perfect information on the other hand, counterfactual deliberation is incorporated in the process of belief updating at unreached nodes, through the selection of the closest deviation-world. Neither of the approaches avoids the ambiguity behind counterfactual reasoning. There is no unique way to choose either the nature of the types and their prior probabilities, or the closest deviation-world.[33] These features will depend on the particular situation in which the game is played.

Regarding the idea that there could be rational or intentional deviations, we face the problem of causality: the rationality of a player depends upon his choices and not vice versa. In other words, a deviation need not be rational because it was chosen by a rational player. Nevertheless, there is some scope for the idea that there could be some kind of law or principle behind people's actions, supported by previous observations, so that after

---

[32] If the players are confortable with the assumption that their opponents could be types who <u>always</u> leave the money.

[33] To be independent of counterfactual reasoning from outside the model, the types should represent a complete description of the game. That is, there can not be mistakes or irrationality or any other explanation beyond the types. Otherwise an altruist could make a mistake and take the money!

observing a deviation we are less willing to give up the assumption of full rationality, including the implementation of actions. However, this criterion *may* yield some un-intuitive results in cases in which the set of parameters in the counterfactual world is not fully described, as in the counterfactual *"Had John jumped off the Empire State building he would have killed himself"* (see section 1.2). Although the issue of whether there was a net seems to be unspecified, Bennett's theory brings it as a necessary feature of the counterfactual world whereas Lewis' does not lead to the same type of revision of the facts holding at the antecedent time.

One could also think of a game where different forms of irrationality and intentions coexist and players have common knowledge of a belief revision policy which, for a given actual world and each possible deviation, pins down the closest counterfactual world.

4) In the introduction we stated that one of the controversial issues regarding the foundations of the backward induction argument in this game, is whether the mere consideration of out of equilibrium situations requires the weakening of the assumption of full rationality in order to avoid a contradiction.

Without any weakening, the term "common knowledge of rationality" is an empty notion. Either mistakes, wrong calculation or wrong beliefs should be introduced to conceive the world of a deviation[34]. However, these forms of irrationality are supposed to occur off-the-equilibrium path and constitute epistemological frameworks within which the deviating behavior can be analyzed. In other words, for the notion of rationality to be meaningful we have to assume that irrational choices are *open* to the players. Players need to have access to these counterfactual scenarios and this access in itself, does not necessarily mean giving up the notion of rationality or the amount of information that players have in the actual world. As it was said before, players will not be rational at all nodes or along all possible paths of the game (otherwise there would be no rational play). They need to be rational in the actual world and in every closest deviation world, for every possible deviation.

The important feature of the present analysis is its capability to deal separately with the different facets of rationality, namely, rationality to choose and implement an action and to form beliefs. It is worth noticing that rationality in choosing an action is the only one that determines node-rationality *before* the players fully follow the consequences of their conjectures. Under Bennett's approach we conclude that common belief in the truth of C_2

[34] Wrong beliefs at off-the-equilibrium nodes reflect some sort of ex post irrationality. Trembles, on the other hand, consist in another form of irrationality ex-post because players that tremble fail to actually perform the right action which is a necessary condition for rationality (see Elster [9] page 13).

and common belief of subgame rationality *ex post* (when the iterative analysis that brings consistency amongst players' beliefs has been performed) are not compatible.

The outcome of this paper should be interpreted in the following way: thinking about counterfactual scenarios à la Lewis provides no consistency problems whereas to do it à la Bennett may render the theory inconsistent, depending on the amount of mutual knowledge or belief that players have. These two cases do not exhaust the possible ways of thinking about counterfactual situations. The contribution of the present work has been to introduce an alternative interpretation capable of showing under which kind of assumptions concerning hypothetical thinking and knowledge we obtain consistent foundations for the backward induction outcome.

References

[1] Aumann, R. "Agreeing to disagree" Annals of Statistics 4 (6) 1236-1239, 1976.

[2] Aumann, R. "Backward Induction and Common Knowledge of Rationality", Games and Economic Behavior 8, 6-19 (1995).

[3] Bennett, J. "Counterfactuals and Temporal Direction", The Philosophical Review 93 (1984), pp. 57-91.

[4] Bicchieri, C. Rationality and Coordination Cambridge University Press 1993.

[5] Bicchieri, C. "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge" Erkenntnis vol 30 (1989).

[6] Binmore, K. "Backward Induction: Reply to Aumann" Working paper, University College London, London WC1E 6BT, UK, 1993.

[7] Binmore, K. "Modeling Rational Players" Part I, Economics and Philosophy 3,1987.

[8] Elster, J. Rational Choice New York university Press, 1986 pp. 13.

[9] Fudenberg, D. and Tirole, J. Game Theory MIT Press, 1991.

[10] Goodman, N. Fact Fiction and Forecast Harvard University Press 1979,1983.

[11] Harper, W. "A sketch of some recent developments in the theory of conditionals" in IFS edited by Harper, W., Stalnaker,R. and Pearce, G. D.Reidel Publishing Company.

[12] Hintikka, J. Knowledge and Belief. An introduction of the logic of the two notions Cornell University Press 1962.

[13] Jackson, F. "A Causal Theory of Counterfactuals", Australian Journal of Philosophy 55 (1977).

[14] Lewis, D. "Counterfactual Dependence and Time's Arrow" Collected Papers II (Oxford, 1986), pp. 32-66.

[15] Lewis, D. "Counterfactuals and Comparative Possibility" in IFS edited by Harper, W., Stalnaker, R. and Pearce, G. D. Reidel Publishing Company, 1981.

[16] McKelvey R. and Palfrey T. "An experimental study of the centipede game" *Econometric*, 60:803-836, 1992.

[17] Reny, P. "Rationality in Extensive-Form Games" Journal of economics perspectives vol 6. no 4 Fall 1992.

[18] Samet, D. "Hypothetical Knowledge and Games with Imperfect Information" Working paper Tel Aviv University December 1993.

[19] Selten, R. and Leopold, U. "Subjunctive Conditionals in Decision and Game Theory" In W.Stegmuller, W.Balzer, and W.Spohn (eds) Philosophy of Economics, Proceedings, Munich, July 1981. page 191-200. Springer-Verlag.

[20] Stalnaker, R. "A Theory of Conditionals" in IFS edited by Harper, W., Stalnaker,R. and Pearce, G. D.Reidel Publishing Company, 1981.