UNIVERSITY OF CALIFORNIA
Los Angeles

The Role of Counterfactuals in the Foundations of Equilibrium Concepts in Game Theory

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Economics

by

Graciela Rodriguez Marine

1995

The role of counterfactuals in the foundations of equilibrium concepts in game theory

Introduction

In spite of the particular conditions under which a collection of strategies is an equilibrium within a game, it is always necessary that players do not find unilateral deviations profitable.[1] In other words, players should not have an incentive to deviate from an equilibrium when their opponents conform to it. On the other hand, players are typically assumed to decide upon *strategies*; that is, they are supposed to choose an action for every possible circumstance in which they might be called to play. This implies that they might have to decide upon their responses at nodes that are not reached under equilibrium.

In simultaneous move games players can not *actually* respond to deviations by their opponents because no player can observe any action before playing. Therefore, hypotheses about events that could have occurred as a consequence of a deviation are constructions that deserve a trivial analysis given that such a dependence does not potentially exist.

The situation is radically different in non simultaneous move games; namely extensive form games with observed actions. Consider for instance an extensive form game with perfect information. Before choosing strategies every player must speculate

---

[1] This necessary condition that all equilibria must fulfill can be represented by the notion of *Nash Equilibrium*. A Nash equilibrium is a collection or profile of strategies, one for each player in the game such that each player's strategy is an optimal response to the other players' strategies.

upon the *possible* outcomes of his play. It could be said that no observation has taken place yet and that therefore the situation is qualitatively similar to that of a simultaneous move game. However, the player who is analyzing the consequences of a deviation ought to consider the possible *reactions* by the players who *observe* his move and play after him.

A premise of the present study is that the reasoning which supports an equilibrium in a simultaneous move game has a different cognitive or epistemological nature compared to that in a non-simultaneous move game. The reason is that only in non-simultaneous move games actions *might in principle* be observed. It can also be claimed that the introspective reasoning accomplished by each player represents these hypothetical thoughts *as if* play was *actually* performed. However, this is true *only* when actions *might potentially* be observed. In this case deviations *might* confer some *information* and therefore have a *causal consequence* upon the play of the other players.

An equilibrium must always be supported by a set of conditionals that describe what would happen in every possible state of the game. These conjectures assure that it is not profitable to deviate and that therefore the corresponding collection of strategies constitutes an equilibrium. These conjectures are of two sorts: the *indicative conditionals* which describe the structure of the game and the features of the equilibrium under consideration and the *counterfactual conditionals* that represent players' hypotheses about what *would have happened had somebody deviated*. The second type of conditionals constitute conjectures about the occurrence of events that are *not expected in equilibrium*, that is, they are conditionals with *false antecedents*.

Let us think of a strategy as a set of maps each going from the set of nodes where this player *might* have to play to the set of actions available to him at that node. Every map represents a possible way in which a player *would have played had that node been reached*. Although the standard models in game theory define strategies in this way, that is, as a set of contingent actions, they do not properly formalize this aspect; hypothetical constructions are implicitly equated with material conditionals.[2] This aspect of the formalization is of crucial importance in terms of the foundation of an equilibrium. For instance, if we assert that the consequent of a material conditional is true then it does not matter whether the antecedent is true or not because the implication obtains in any case. Consider the following conditional: *"had node 'n' been reached then action 'a' would have been played"*. If we fix the action taken at that node by means of introducing a behavioral assumption and analyze the conditional as a material one, then the truth of the predicate *"node 'n' is reached"* is irrelevant. The conditional would be true regardless of whether the node is actually reached. Off-the-equilibrium contingencies *have* a false antecedent and therefore, the truth of the consequent can not affect that of the conditional if taken as a material one.

Out-of-equilibrium conditionals are important because a player can not *choose* a strategy if he can not *assert* what *would have happened* otherwise. In other words, the play of a given equilibrium by each player is justified if and only if each of them either knows or believes that *"had he deviated he would have not been better off."* Our

---

[2] A material conditional is a conditional whose truth depends upon the truth of its antecedent and consequent (conditionals in mathematics always satisfy this criteria). The material conditional "if P then Q" can be defined as "either Q is true or P is false" or "it is not true that P is true and Q is not".

main premise is that in order to obtain a proper foundation for the notion of equilibrium in extensive form games players need to assert the truth condition of these counterfactuals in an appropriate framework.

Within the theory of games it is typically assumed that players construct hypotheses concerning the contingent play of their opponents based upon the assumption that it is common knowledge that everyone aims at maximizing his payoffs. In the present study we assume in addition that players are endowed with a common framework to judge the truth of the counterfactuals which support the equilibrium under consideration. With this purpose we apply two closely related theories of counterfactuals: those developed by David Lewis and Jonathan Bennett respectively.[3] Our aim is to study the consequences of respectively incorporating these accepted frameworks to the foundations of two refinements of the basic concept of Nash equilibrium in extensive form games: backwards induction and sequential equilibrium. We test the consistency of the results implied by these expanded frameworks and analyze the outcomes of having players form beliefs in the manner prescribed by our interpretation of these theories of counterfactuals.

The monography is organized as follows. In the first chapter we address the issue of whether common knowledge of rationality leads to backwards induction. This question has promoted a considerable yet unsolved controversy in the literature. Our aim is to specify the conditions under which the results achieved in the literature obtain in terms of our framework. In the second chapter we analyze two refinements to the notion of sequential equilibrium in signaling games: the *Intuitive Criterion* postulated

---

[3] These theories are explained in detail in sections 3.2 and 3.3 respectively.

by Cho and Kreps and *Divinity* postulated by Banks and Sobel. Within our interpretation of the theories of counterfactuals presented in the first chapter we analyze the extent to which a player might signal a piece of information that his opponent lacks. Based upon this analysis we impose different restrictions upon the beliefs of the uninformed player and establish the conditions for existence of different sorts of equilibria. In addition we propose a modification to *Divinity* in order to further refine the set of equilibrium outcomes.

I. The backwards induction solution to the centipede game

1. Introduction

As it was pointed out in the general introduction, players' reasoning about the responses to their actions are crucial constituents of an equilibrium. As it is extensively acknowledge in the literature, these thought experiments depend not only upon the logical framework used by the players, but also upon their knowledge and beliefs about the game and about the knowledge and beliefs of the other players.

In extensive form games with perfect information, the backwards induction refinement requires that each action that forms part of an equilibrium strategy be a best response in every possible subgame starting from every final node of the game. Players must hypothesize about the outcome at *every possible subgame* regardless of the identity of the player at each node and by a backwards tracking reasoning, completely construct their strategies as they reach the first node at which they might have the chance to play.

It is typically asserted that in the case of games with perfect information, common knowledge of maximizing behavior leads to the backwards induction equilibrium because it guarantees that players have no incentive to deviate along any possible path of the game (see Aumann [1]); yet this is an open matter. As it has been asserted before, in these types of games deviations *might* have an informational value and *might* lead to responses. Therefore, every player needs a conjecture not only about the corresponding counterfactual scenarios at which each player might find himself but also about the conjectures of his opponents regarding his conjectures and so on.

In the literature of non cooperative extensive form games with perfect information, the centipede game is one whose backwards induction solution still

motivates a considerable amount of disagreement concerning its logical foundations. Two issues sustain the controversy regarding this game. 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 backwards 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 [6] 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 issue, assume that it is possible for the players to have common knowledge of rationality. The issue of how to derive the backwards induction outcome when hypothetical thinking is present is also a matter of controversy. Binmore [6] proposes to enlarge the model to allow impossible events to occur. He does this by assuming a game with an infinite set of players so that there exists a non-empty set of measure zero of irrational players. On the other hand, 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 backwards 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 that common hypothesis of rationality at each node implies backwards 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 chapter is to test the internal consistency of the solution concept by presenting a formalization of the backwards induction solution to the centipede game capable of incorporating counterfactual reasoning at off the equilibrium nodes. The aim is to find sufficient and necessary conditions regarding players' knowledge and beliefs capable of yielding the truth of the supporting counterfactuals in order to justify this equilibrium concept.

With the purpose of formalizing counterfactual reasoning two closely related theories of counterfactuals are introduced. These theories have been proposed by David Lewis [15] and Jonathan Bennett [3] respectively. Under our interpretation of Lewis's approach and the assumption of common knowledge of rationality (as it will be defined below) the backwards induction outcome can be obtained. The reason is that in this case, common knowledge of rationality can be expected to hold after a deviation given that, within our interpretation of this theory, players are not led to update their beliefs concerning the rationality of their opponents at counterfactual scenarios.

Under our interpretation of Bennett's theory 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, backwards induction can not be supported. Relaxing the assumption of common *knowledge* or rationality in favor of common *belief* implies that there may exist scenarios compatible with backwards 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 chapter is as follows. The first section explains the notion of a counterfactual and the nature of the counterfactuals involved in games. The second presents the framework and formalization of the backwards induction solution in terms of counterfactual reasoning. The third section incorporates the two mentioned theories of counterfactuals to analyze the truth conditions of the corresponding conditionals under different assumptions of rationality. 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.[4] 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

---

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

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. Moreover, Stalnaker [20] observes that "the falsity of the antecedent is never sufficient reason to affirm a conditional, even an indicative conditional." That is, conditionals no matter whether indicative or subjunctive establish a connection or function between propositions that is not necessarily represented by this 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 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. The conditional will be rendered true by just taking care of the truth condition of the corresponding propositions when we know that the reason why John passed was not that he studied.

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 [11]). 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[5]. 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

---

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

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.

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. Moreover, there are cases in which the nature of the connection can not be completely established from observation. 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 or capacity. 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.

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 the solution 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 framework (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 ([15]&[16]) have suggested two closely related theories of counterfactuals (see Harper [12]).

When we believe that the antecedent is false (for instance, when the antecedent entails a deviation by some player) the thought experiment within which the antecedent is true may not consist in 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 *may 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 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 should be commonly held for the equilibrium outcome to be consistent.

Possible world theories establish a *metric* to evaluate which of the possible explanations should 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 criteria to select which of the above worlds is the most plausible. In the case of game theory, this criteria requires a behavioral assumption that in general is represented by the concept of rationality. In other words, we need to find the world (or world*s*) 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...[6]

Rationality in strategic contexts is a complex phenomena. 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 in their *decisions* about how to further play the game. If a deviation is understood as some 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

---

[6] This example is discussed in Jackson [12] and Bennett [2].

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 interfere with the outcome of the fall. The question is how valid is this criteria within a game because the counterfactual consequent, that is the response to a deviation, may change if the occurrence of a deviation implies further occurrences of other deviations.

On the other hand, there is the issue of what are the parameters or features of the world that we are allowed to change. 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? Obviously no. 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?

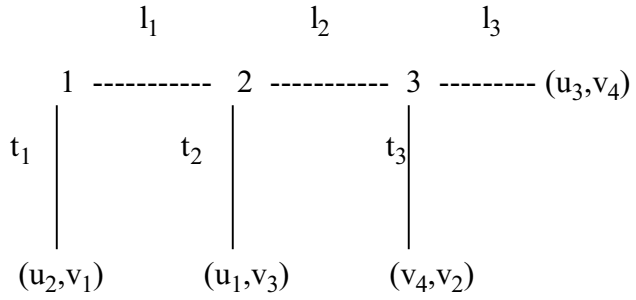
In a world in which John jumps his friend gets some information that makes him change his decision (we assume he would have not jumped in the absence of a net). However John's friend could have updated his stock of beliefs to attribute the hypothetical occurrence of the jump to some unexplainable reason 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 backwards 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 equal to u_2 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_3 whereas player 1 gets u_1. If player 2 leaves the money then player 1 has the final move. It he takes player 1 and player 2 get respectively u_4 and v_2. Otherwise they get payoffs equal to u_3 and v_4 respectively.



The pair of letters between parenthesis at the termination points represent the players' payoffs and they are such that $u_2 > u_1$, $u_4 > u_3$ and $v_3 > v_2$. ('u' and 'v' stand for player 1's and 2's payoffs respectively).

The numbers between parenthesis represent the label 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 backwards 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 model 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. In a backwards induction solution, beliefs are not updated as the play proceeds from the beginning and this implies that the hypothetical play of l_1 cannot have any repercussion upon further decisions. 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 to deviations.

The centipede game is a game of perfect information. This implies that, as the game is played, players' expectations about future play *could* either be confirmed or deceived yet, this has no role in the backwards argument. Such is the nature of backwards as opposed to forwards solutions. Backwards tracking arguments do not allow beliefs to be updated because every future contingency *has been already evaluated and discounted all the way up to the root*. Forwards tracking solutions, on the other hand, bear the time-inconsistency problem very well known in the literature. For instance, the case of a bubble: assume that players start building the expectation that as they go along the centipede the money will be left on the table in the next

round. It is clearly inconsistent to do so at the previous to last node unless the player who moves there believes that the last player is irrational in a very rudimentary way[7].

The previous perspective seems to leave us with a trade off between both equilibrium concepts. However, what the backwards induction solution needs to be complete is a theory which specifies how to reason at counterfactual scenarios. In the past years some agreement concerning the role of counterfactual scenarios has emerged within the literature (see Binmore [7], Bicchieri [5], Samet [18]). Even Aumann [1] who proves that common knowledge of rationality implies backwards 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" (op.cit page 17). He asserts that the "if ...then" clauses involved in equilibrium are not material conditionals (as in mathematics) but substantive conditionals.[8]

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

---

[7] That is, he is incapable of choosing between two different certain payoffs.

[8] 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.

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$\subset$**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} , $a_{1n} = t_n, l_n$ } n=1,3 ;

A_{2n} = { a_{2n} , $a_{2n} = t_n, l_n$ } 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.[9] In addition let h^{n+1} include the action taken at node n:

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

Given that this is a game with perfect information, h^n represents players' knowledge about the past play which lead 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={$P(z_1)$;$P(z_2)$,$P(z_3)$,$P(z_4)$}

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

---

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

(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:

$S_1=\{s_1, s_1 = t_1t_3 , t_1l_3 , l_1t_3 , l_1l_3 . \}$

$S_2=\{s_2, s_2 = t_2, l_2 . \}$

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. $U_i :H \rightarrow \mathbf{R}$ 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)" "(pvq)" and "($p \Box\!\!\rightarrow q$)";[10]

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"

"~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.

---

[10] 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.

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 a 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. Needless to say that 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 presented in Hintikka [13].

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 he knows far enough and 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 $\text{"}\sim p\text{"} \in \mu$.

That is, a model set can not have as members a proposition together with its negation.

(C.&) If $\text{"}p\&q\text{"} \in \mu$, 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 $\text{"}pvq\text{"} \in \mu$, 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 $\text{"}\sim\sim p\text{"} \in \mu$, 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 $\text{"}\sim(p\&q)\text{"} \in \mu$, then $\text{"}\sim p\text{"} \in \mu$ or $\text{"}\sim q\text{"} \in \mu$ (or both).

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

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, $"P_i\, p"$ ∈ μ. The substance of the statement $"P_i\, p"$ can not be given a proper meaning unless there exist 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 $"P_i\, p"$ ∈ 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 *a* such that *p* ∈ μ*.

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 $''K_i\ p''\ \in \mu$, then $p\ \in \mu$.

This says that knowledge cannot be wrong. In other words, if *i* knows that p then p is true.

(C.~K) If $''\sim K_i\ p''\ \in \mu$, then $''P_i \sim p''\ \in \mu$.

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 $''\sim P_i\ p''\ \in \mu$, then $''K_i \sim p''\ \in \mu$.

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 <u>*any*</u> 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:

μ={ $''K_i\, p''$, p, $''P_i\, u''$ }

μ*= { $''K_i\, p''$, 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 [13] 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[11] 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*):

---

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

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

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 <u>*indefensible*</u> 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 viceversa 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 viceversa (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 an 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.[12]

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: existency: 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 beliefs 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

---

[12] Remember that '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.

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 to note that within this framework to believe that one believes something does not imply that one believes it. However if the mutual belief operator is defined as the conjunction of the different degrees of knowledge then we can obtain implications of the following form: if everybody believes that everybody believes then, everybody believes.

2.4 The backwards induction solution

Before stating the definition of the backwards induction equilibrium we need to introduce the following concepts:

1) Let $G(h^{n'})$ denote the game that given the public history of the game 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).

Now, a backwards 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 backwards 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]).

One of these two Nash equilibria satisfies this requirement and therefore constitutes the backwards induction solution:(t_1t_3,t_2).

The standard argument for the backwards induction solution in this game 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.

In general the justification goes as follows: 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 backwards induction equilibrium by considering strategies as contingent events and then introducing a theory to solve these counterfactuals. *Within each of the theories* 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 beliefs that players commonly hold at all possible nodes.

3. The backwards induction solution and the theories of counterfactuals

In this section two different theories of counterfactuals are applied to analyze the backwards induction outcome.[13] Before doing so, a few preliminary issues should be addressed.

Under equilibrium, the *factual* or *actual* world will be defined as the world in which the equilibrium strategies are contingently chosen. That is, a world in which counterfactuals C_3 and C_2 are true. However, as it was already stated, even within the actual world, some equilibrium strategies *might* not be actually played. For instance,

---

[13] For a detailed presentation see Lewis [13] (alternatively Lewis [7] (pages 57-85)) and Bennett [2].

player 2 does not have the chance to play under any of the two Nash equilibria of the centipede game. Therefore, 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 occurrence. That is, we have to answer the following questions: 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"?

Within the present framework, players can have no knowledge regarding their opponents' off-the-equilibrium-path play because no observation takes place at those nodes. Therefore, players can only have *conjectures* or *beliefs* regarding the truth of these events. For instance, a player might know his own decisions off-the-equilibrium path, but he cannot know his opponents' play at nodes that are not reached under equilibrium.[14]

Regarding the second question, there is a crucial concept whose epistemological status must be defined in order to assert the truth of the mutual conjectures concerning off-the-equilibrium path play. More specifically, the question we need to consider is whether players can *know* that their opponents are rational or have any other type of behavior. Before addressing this matter, a few definitions need to be introduced.

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

---

[14] This is a non cooperative game where no communication takes place.

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 behaviour subject to exogenous beliefs. This will be defined as rationality *ex ante* to stress the idea that beliefs need not be correct. On the other hand, rationality *ex post* will be considered to incorporate in addition 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.

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 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.[15] 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. However on the other hand, a player may play in a way in which his opponent defines as "rational" by pure error and therefore,

---

[15] "*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.

observation would not necessarily provide enough information to establish this type of knowledge even under full observability. It is clear at this point that either a bayesian 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* might be justified. One could think that players *might* know that their opponents maximize given belief 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.[16] *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.

---

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

In addition we need to consider that information might 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 *es ante* rational. However, if there are 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 might 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 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 an exposition of the results obtained under two theories of counterfactuals. Afterwards a formalization of these results within Hintikka's semantical system will be presented.

3.2 Lewis's theory of counterfactuals

Lewis's theory is based on two fundamental concepts: i) the asymmetric openness of time and ii) the notion of possible worlds.[17]

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's 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; an scenario that despite his 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).

Lewis assumes that there exist 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

---

[17] For more detailed exposition see [6], [7] or the appendix.

whatever this 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, call it, w_1, is in matters of facts similar to the actual world, call it, w_0, until shortly before the deviation is supposed to obtained. At the antecedent time (call it t_P) "the deterministic laws of w_0 are violated at w_1 in some simple, localized, inconspicuous way. A tiny miracle takes place." (Lewis [15] page 44). At w_1, a mistake produces the corresponding miraculous deviation (the occurrence of the deviation does not necessarily imply that the player chose to deviate. It only implies that he *did* it). No further "miracles" occur and after t_P, w_0 and w_1 diverge in matters of facts.

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's 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 amongst 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's 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" (op.cit page 45). In Lewis's theory worlds would be ranked from the closest to the farthest in the following: 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's bias towards the importance of facts previous to t_P allows to fix 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. [18]We first need to search for the closest hypothetical world where a deviation occurs. Within Lewis's paradigm, the smallest miracle that can produce a "3"-world, 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 behaviour in the actual world as close as possible*. Under any definition of rationality[19] 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's metric the closest to the equilibrium world w_0 what allows 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), that is, they make mistakes intentionally. In this case player 1 would play l_3 so that the counterfactual C_3 is false.[20]

---

[18] 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.

[19] 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.

[20] $(P \mathbin{\Box\!\!\rightarrow} Q)$ is false iff $(P \mathbin{\Box\!\!\rightarrow} \sim Q)$ is true. See [7].

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. However, is in the interest of the present work to discuss whether *in the absence of trembles* as they are defined in the literature, the typically assumed "common knowledge of rationality" is sufficient for the backwards induction outcome. 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.[21]

Under our interpretation of Lewis's theory, deviations are not incompatible with players' *ex post* rationality because there is no update of rationality after a deviation. Deviations are incompatible with *full* rationality. There is, nevertheless, 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[22]. However, this does not hold at any of the other nodes. At node 2 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 assumed at this stage.[23] Under this

---

[21] 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.

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

[23] At node 3 any assumption of rationality without miracles or mistakes would bring "t_3" true.

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 basis of the *common hypothesis 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.[24]

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 tremble 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.

---

[24] 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.

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?

Let us for expositional purposes 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	t_2,t_3-world
$w_{R'}$	subgame rational with wrong beliefs	subgame rational	t_2,t_3-world
$w_{R''}$	subgame rational with right beliefs	subgame irrational	l_2,t_3-world
w_I	subgame irrational	subgame rational	l_2,l_3-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. In this way we rule out w_I as a possible world. Now we have to find the closest deviation-world and see whether "t_2" is true in it.[25]

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 $l_1t_2t_3$-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

---

[25] 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.

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.[26] Therefore, we are left with w_R where player 2 is supposed to play t_2 given the assumption of node rationality. [27] In this case we obtain the backwards 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 when the players are rational and that this may disturb the previous ranking.[28] However even if we considered 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 allowed* 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 backwards induction. The key element in this argument is that beliefs are "revised" in such a way that *common hypothesis of node rationality* is still possible *after a deviation*.

---

[26] The purpose is to avoid back tracking arguments. See the appendix and Jackson [12].

[27] It could be said as it is implied in Binmore [2] 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 rationality without trembles.

[28] This is one of Binmore's remarks. See [2]

Given our definitions, *ex ante* rationality can be mutually known amongst 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 hypothetical play.

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

Definition: It is said that ($P \mathrel{\square\!\!\rightarrow} Q$) 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

backwards directions. If Q is true in each of these worlds then the counterfactual is true. In other words, once the deviation occurred, we reason backwards 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 endogenizes beliefs and hence rationality 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 theory 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 theory under the assumption of rationality without trembles leads 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. 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. Within Bennett's worlds deviations are intentional, that is, players are supposed to be aware that they are deviating.

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*.[29] Therefore, if player 1 is supposed to be *es 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. [30]

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*

---

[29] Miraculous mistakes as well as standard trembles are ruled out by assumption. Moreover node-rationality is compatible with wrong beliefs.

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

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. Under the assumption of common belief of subgame rationality and the theory under consideration, players will not only face an inconsistency. This inconsistency will be commonly believed. The key feature that brings this result is the combination of the assumption that previous necessary play carries *knowledgeable consequences* and the supposition that players have *common* belief in subgame rationality.

3.4 The formalization of the results.

This section analyzes the conditions under which the backwards induction outcome obtains in the presence of different levels of mutual knowledge and belief within the semantical system presented in section 2. Therefore, this section enlarges the results already presented not only by explicitly modeling them but also by offering a richer range of outcomes.

The following definitions establish within our language the concept of node rationality ex ante under fully intentional play. The axioms, on the other hand, state structure of the game and the players' rules of inference.

Definitions and axioms:

(A1) Structure of the game: payoffs, available actions and order in which players move.[31]

[31] This was presented in section 2.

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 counterfactual three is true under any theory of 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) exhaust all the possible causes of a deviation and define strategies as contingent constructions. Mistakes or any alternative explanation (for instance imperfect information about the payoffs) could have been allowed here as another proposition in the disjunction. After presenting the results under full intentionality we will introduce this option.[32]

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:*[33]

$h^{j} = \{"K_2\, l_1"\,, "B_2 \sim R_{11}"\}$; $h^{j} \subset \mu_2{}^{j}$ (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)[34]

---

[32] Recall that under our interpretation of Lewis's 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.

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

$h^{iij} = \{"K_2\, l_1"\,, "B_2\,(R_{11} \& B_1\, B_2 t_3\, \&\, B_1\, l_2\,)"\}$; $h^{iij} \subset \mu_2^{iij}$ (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 one would play t_3 in every possible world in which he is node rational at this last node. These worlds are

---

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

considered closer to the actual where he is rational. Hence any theory of counterfactuals would render this counterfactual true.

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 exist a defensible set that represents the player's states of mind that is compatible with the truth of the corresponding counterfactuals so that the backwards induction outcome is chosen. 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.[35]

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 $(B_2 t_3)]$" $\in \lambda$ for i=1,2 that is that "K_i [ck $(B_2 t_3)]$" 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

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

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 exist 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 exist 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

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 one believes that player two 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 exist 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 exist 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 [3] we reduce the degree of mutual knowledge to find the amount of knowledge that is necessary to guarantee backwards induction.

Case 2:

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

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

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 backwards induction play. In words, 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 persist 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 exist 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}&~$K_2K_1R_{22}$)"*}; $h_1^{jv} \subset \mu_1^j$. 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 exist $h_1^{jv} \subset \mu_1^j$ where μ_1^j is an alternative world to μ_1 with respect to player 1 and $\mu_1^j \subset \lambda_1$.

<u>Case 3</u>:

E_1 = {*"K_1 (A10),"$K_1(R_1)$","K_1 R_2","K_1K_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_2K_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. So there is a defensible set representing a world alternative to the actual with respect to player 1 where the contingent play required for backwards 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 criteria to play. Backwards induction can not be supported in this case and for any higher degree of knowledge.

Before dropping the assumption of knowledge of rationality it is worth noticing that had Lewis's theory as we interpreted it been modeled, there would have been some alternative in the disjunction of axioms (A4) and (A5) such that the deviation need not be intentional. Moreover (A11) should be replaced by the alternative ranking of scenarios under this theory. In Lewis's theory, worlds in which some departure from perfect performance explains the deviation are the closest ones. In

this case players knowledge and beliefs 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 backwards 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.

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 exist $\mu_i{}^j$ 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 $h^{jj} = \{"K_2\, l_1"\ , "B_2\ (R_{11} \&\ B_1\ B_2 l_3\ \&\ B_1\ R_2)"\}$. Due to *"ck(A11)"*, it is common knowledge that h^{jj} is included in a possible state of affairs $\mu_i{}^{jj}$ 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*) there exist $\mu_i{}^{jj}$ such that $"K_2\, l_1\ \& B_2\ (R_{11} \&\ B_1 B_2 l_3\ \&\ B_1\ R_2)" \in \mu_i{}^{jj}$ where $\mu_1{}^{jj}$ and $\mu_2{}^{jj}$ 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 one believes that player two 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 μ_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 exist μ_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 μ_1^{iii} and μ_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 μ_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 backwards 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 the 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 backwards induction outcome and where common belief in node rationality can be assumed*[36]. 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.

4. Concluding remarks

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

i) Players do not update their assumption regarding rationality when they observe a deviation and we assume *common knowledge* (or belief) of rationality. This result can be obtained under our interpretation of Lewis's theory of counterfactuals.

Within Lewis's theory there are worlds in which players deviate and are 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

---

[36] Actually it is easy to prove that a smaller degree of mutual belief is necessary and sufficient. In the present game, player one needs to believe that player 2 is rational and that player 2 believes that player 1 is rational.

knowledge of *ex ante* rationality the former type of worlds are closer to the actual than the latter. In the counterfactual world where a deviation occurs the *a priori* belief that the player was not going to deviate at the node where the deviation occurred need to be given up. However under the proposed metric this does not lead to reject the belief or knowledge in further node rationality. The reason lies in the way in which beliefs are updated. More precisely, deviations need not have *causal* consequences. A deviation occurs but no intentionality underlies it. Deviations might be related to the performance of the action and not with the decision itself. Therefore rationality in the process of decision 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 deviators (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 the present work this obtains when player 2 knows that player one 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 would mean that if the players face the game again with one more node the theory will 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]).

The truth of the counterfactuals required to support backwards induction solution 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 leg 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 backwards 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 backwards 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 naturally infinity. In this case, common belief in node rationality can be *only*

preserved within the players' equilibrium worlds (which means that backwards induction can be obtained). Any world in which a deviation occurs will not be consistent with the assumption of common belief in node rationality *ex post*. In other words common belief in node rationality needs to be dropped in alternative-out-of-the equilibrium worlds.

The previous results lead us to the following conclusions:

1) Either we allow for non updating and keep full knowledge or allow for updating and relax the degree of mutual knowledge. Another option is to substitute "knowledge" in favor of "belief".

2) 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. The access in itself does not necessarily mean giving up the notion of rationality or the amount of information that players have. We need to *assume* what the occurrence of a deviation means in terms of belief updating. The theories of counterfactuals above presented reveal that *there is no unique way* to solve the context dependence in which counterfactuals are generally stated.

3) 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?

Given a game and an environment we assume that players have been already provided with a theory to form beliefs which is common knowledge or at least common belief. However on the other hand, we can also assert that rational players will not follow irrational theories.

The answer to this matter crucially depends on how we think about rationality. 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 deviation. If rationality is seen as a set of rules, then Bennett seems to be a more reasonable analysis. However there is no clear answer or consensus at this respect. The drawback is that Bennett *might* yield some un-intuitive results in cases in which the set of parameters is not fully described in the counterfactual world as in the counterfactual *Had John jumped off the Empire State building he would have killed himself* (see page 7 above). 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's does not lead to the same type of revision of the facts holding at the antecedent time.

One could also think of a meta game where players choose a theory of counterfactuals. For example we can allow mistakes and intentions to coexists. Nevertheless, players need to have a common criteria to decide upon the metric concerning the similarity of the possible worlds. It might be suggested that the criteria for the metric could depend on the stage of the game yet it is worth to notice that this rule should be commonly held.

4) According to our interpretation of Lewis's 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 backwards induction result is that it excludes further mistakes or trembles. It is worth to remember that Lewis's imposes this condition to guarantee the closest resemblance to the factual world.

5) Under our interpretation of Bennett's framework on the other hand, deviations give some information about the beliefs of the deviating players. 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 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 to the worst payoff with certainty. The important feature of the present analysis is its capability to deal separately with the different facets of rationality, namely rationality in action and in beliefs. It is worth noticing that rationality in 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 and common belief of subgame rationality (node-rationality *at all* nodes) *ex post* (when the iterative analysis that brings consistency amongst players' beliefs has been performed) are not compatible.

6) There is a second interesting issue concerning the analysis of counterfactual scenarios in game theory. It seems that the consideration of out of equilibrium situations that is needed to justify an equilibrium requires the *weakening of the full rationality* assumption at least in the hypothetical worlds if we want to avoid a contradiction. Without any weakening the term "common knowledge of rationality" is an empty notion in an extensive form game. Either mistakes, or wrong hypotheses

should be introduced in counterfactual worlds. These mistakes or wrong beliefs are supposed to occur off the equilibrium path and they constitute epistemological frameworks within which deviating behavior can be analyzed. Wrong beliefs at off the equilibrium nodes reflect some lack of successful iteration along that path *ex post*. Trembles, on the other hand, consist in another form of irrationality (ex-post-decision) because players that tremble fail to actually *perform* the right action which is a necessary condition for rationality (see Elster [9] page 13).

In the original backwards induction argument, deviations do not reveal meaningful consequences. This is what happens within our view of Lewis's miraculous worlds. The way the hypothetical scenario of a deviation is brought about is irrelevant to the assumption about further rationality. Other analyses consider that deviations provide some meaningful information to the players (see Binmore [4]). Deviations may lead to further deviations, either in a correlated or uncorrelated way, or they may reveal something about the beliefs of the player that has deviated. Bennett's framework is an example of this last type of interpretation.

The outcome of this paper should be interpreted in the following way: thinking about counterfactual scenarios à la Lewis provides no consistency problems, where to do it à la Bennett might make 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 backwards induction outcome.

II. Belief updating and equilibrium refinements in signaling games

1. Introduction

Many issues modeled in economics are concerned with non-cooperative scenarios in which there is one agent who is privately informed about the state of nature and another who is uninformed. The informed agent takes an action which is observed by the uninformed agent who, after drawing certain inferences, takes an action in response. The payoffs to both parties depend on the state of nature and the actions taken. A signaling game is a stylization of this type of scenario: by taking an action, the informed agent sends a signal or message to the uninformed who, based upon this observation, constructs beliefs regarding the true state of nature which is unknown to him. The uninformed player responds by taking an action which is a best response to these beliefs. The message sent by the informed player maximizes his expected payoff given his beliefs concerning the response of the uninformed player.

A Nash equilibrium of this type of game is a profile of behavioral strategies such that each of them constitutes a best reply with respect to the other. This means that no player can individually gain by deviating from a Nash equilibrium when the other player plays in accordance with it. However, in order to decide how to play, the informed party needs to consider the likelihood with which player 2 might respond to his different messages. The only requirement that Nash equilibrium imposes is that the equilibrium response by player 2 to unsent messages deters the informed party from abandoning the strategy prescribed by the equilibrium.

The concept of Nash equilibrium requires that each player chooses a strategy which maximizes his expected payoff assuming that the other players play *in accordance* with the equilibrium. For this reason it *may* prescribe responses that are

not optimal in the face of a deviation. When this occurs, a player might profit from playing a non-equilibrium strategy in response to a deviation (note that this might trigger further deviations). Van Damme ([22] page 10) characterizes this type of equilibria as not *self-enforcing*.

It is accepted that, in order to be self-enforcing in the sense just described, an equilibrium must recommend maximizing behavior at *every* possible situation at which a player might have the chance to play. In the case perfect information games, the concept of subgame perfection is sufficient to fulfill this requirement. However, this is not necessarily the case with an imperfect information game; namely when there is at least one move by a player who does not know which action preceded his turn. Among the solutions proposed to guarantee maximizing behavior at all nodes while encompassing scenarios in which a player might not know with certainty which action was played before, there is the notion of sequential equilibrium.

A Nash equilibrium is sequential *if* there exists a probability distribution over the states of nature such that the uninformed agent maximizes his expected payoff in the face of a deviation. In other words, under the requirements of sequential equilibrium responses to *unsent* messages should *also* be best replies based upon consistent beliefs. Players use Bayes' rule to compute their beliefs regarding the states of nature along the equilibrium path. By this rule the uninformed party can calculate the likelihood of every state of nature conditional on every message that is sent under equilibrium. In addition it is assumed that off-the-equilibrium messages constitute zero probability events. Therefore, given that Bayes' rule is not defined for conditioning events of this sort, beliefs based upon off-the-equilibrium messages are left undetermined.

The notion of sequential equilibrium only imposes an *existence* requirement upon the set of beliefs that support players' best responses at nodes which are off the equilibrium path. These beliefs do not need to satisfy any further requisite neither is there a rule to compute them. As a consequence, many equilibria might remain even after requiring optimal responses at every information set.

In terms of signaling games, sequentiality imposes no restriction upon the interpretation of messages which are sent off-the-equilibrium path. For instance a player might believe that had his node been reached his opponent would have played a strictly dominated strategy. This might happen because beliefs formed after deviations are not constructed upon the assumption that they signal intentions on the part of the informed player and therefore are considerably unrestricted.

With the purpose of further refining the concept of sequential equilibria several equilibrium notions have been introduced in the literature. Among them are Cho and Kreps' *Intuitive Criterion* [8] and Banks and Sobel's *Divinity* and *Universal Divinity* [2]. The aim of these notions is to restrict possible inferences at information sets which are off-the-equilibrium path by eliminating those beliefs which do not survive some stylization of the hypothesis that deviations could be intentional.

In this chapter we explore the consequences that alternative ways of drawing inferences from deviations have upon the existence of Nash equilibrium and the mentioned refinements to sequential equilibrium.

Consider an equilibrium under which regardless of the state of nature the informed party always sends the same message; this called a pooling equilibrium. Given the distribution from which nature is drawing its states, the uninformed player calculates by Bayes' rule the probability that he faces each state conditional on the

equilibrium message that he has received. In this case the conditional posterior of each state given the equilibrium message is equal to the prior probability of the state. Therefore, under equilibrium no new information is revealed; the uninformed player follows the prior distribution over the states of nature. Now assume that a deviation occurred. This occurrence does not intrinsically confer any information unless we define some concept of rationality together with a theory of how to interpret deviations. Assume that players are rational as it is typically argued in the literature. The question that remains is how to model this deviation. Applying the framework of possible worlds introduced in the previous chapter we have to find the smallest departure from the equilibrium under analysis that will bring the counterfactual world of a deviation. If players are rational at least two possible interpretations seem feasible: 1) deviations are mere hypothetical constructions; that is, they are not intentional and therefore can not reveal further information and 2) deviations are intentional and should be analyzed as the outcome of a rational decision process.

Consider the first case. If a deviation does not confer a signal, one option for player 2, given that he is rational as well, is to adopt the prior distribution which is assumed to be common knowledge as his beliefs regarding the likelihood of the states of nature. However, if beliefs at nodes off-the-equilibrium path are computed in this way existence of Nash equilibrium can not be guaranteed: the receiver's response might not deter the sender from sending a non equilibrium message.

In the light of this outcome two questions can be posed: i) why should a deviation be thought of as unintentional? and ii) why should the second player use the prior distribution as his beliefs off-the-equilibrium path in this case?

With respect to the first question we consider that a deviation might be rationally chosen, and therefore intentional, *only if* it is expected to be profitable. However, by definition, deviations from equilibrium might be profitable *only when* they lead to further deviations. In addition these "further deviations" also need to be compatible with the assumption of rationality and fully intentional behavior. As it was already illustrated in the previous chapter not every game can consistently encompass the compatibility of the assumption of common knowledge of rationality and fully intentional behavior. Therefore the answer to this question crucially depends on the structure of the game under consideration.

With respect to the second question, the answer can only be solved within an empirical context. A player might think that in the face of unreliable inferences he might trust a more genuine piece of information; namely, the prior distribution over types. However, there is no unique way to address this issue. Alternatively, one can also pose the following question: why should a player revise his prior probabilities which are common knowledge conditional on the occurrence of an event that is not supposed to confer information?

Consider now the case in which deviations are intentional and players rational. The goal in this case is to refine the set of posterior beliefs of the player who faces a deviation under the assumption that the corresponding deviation might provide additional information about the state of nature. In terms of the literature, both the *Intuitive Criterion* and *Divinity* propose a framework to restrict beliefs regarding the likelihood of the states of nature in situations in which deviations can be rationally explained.

Roughly speaking, Cho and Kreps' *Intuitive Criterion* proposes that states of nature which can not lead to a profitable deviation by the informed player be eliminated from the game or attached zero probability by the uninformed player. In addition, Banks and Sobel assume that the posterior conditional probabilities of those states of nature which, relative to others, are less likely to lead to an intentional deviation be revised downwards with respect to their prior unconditional probabilities. They further propose an iterative procedure, whose equilibrium outcome they call *Divine,* based upon this refinement.

Both the *Intuitive Criterion* and *Divinity* analyze the intentionality of deviating play while taking for granted that equilibrium responses would have occurred in response to equilibrium play. Although consistent with the features of the equilibrium world that one might want to preserve, this assumption has been challenged by Van Damme ([21] page 281) and it is discussed in depth in section 2.

Regarding the *Intuitive Criterion* it is not clear why players should be completely eliminated from the game given that the sequential reasoning starts with a tentative hypothesis that can lead to further reactions and reconsiderations of the initial assumption. This issue has been pointed out by Van Damme as a drawback of this method ([21] page 281-282). On the other hand this test does not impose any restriction upon the set of sequential equilibria when all the types could to -some extent- potentially benefit from a deviation.

Regarding *Divinity,* when the equilibrium payoffs for every sender dominate any alternative payoff in the game *any* conjecture by the receiver supports the equilibrium. In this case Divinity does not refine the set of sequential equilibria. Banks and Sobel justify this feature of their proposal by explaining that in this case the

receiver should be *truly surprised*. However if this is the case then a deviation should be considered meaningless and the comment provided in the previous case holds. Moreover there is another circumstance in which *Divinity* does not refine the set of sequential equilibrium outcomes. This occurs when the issue of whether player 1 might benefit from a deviation depends only upon the reply by player 2 and not upon his type. However a deviation in this case should also be consider as *truly surprising* in Banks and Sobel's terminology. In section 4 we introduce a modification to Divinity based upon the assumption that whenever the message of the informed player does not convey a signal, the uninformed player uses the prior distribution over the states of nature. With this modification the set of sequential equilibria can be further refined although existency is not longer guaranteed.

This chapter is organized as follows. Section 2.1 presents a formalization of a signaling game and introduces the concept of Nash and Sequential equilibria. Section 2.2 the beer-quiche game from Cho and Kreps [8] together with their *Intuitive Criterion*. Section 2.3 presents Banks and Sobel's concept of *Divine equilibria* and *Universally divine equilibria* together with some examples that illustrate them. Section 2.4. compares the concepts presented in its previous two sections. Section 3 presents an alternative criterion to restrict beliefs at information sets that are off-the-equilibrium path based upon the interpretation of Lewis's theory of counterfactuals which is presented in the previous chapter; namely that deviations do not lead to further updating of beliefs. We assume that the uninformed player uses as his beliefs the prior distribution over types given that deviations do not confer new information. In this case Nash Equilibrium need not exist as the beer-quiche game presented in section 2 illustrates. In this section we characterize the situations in which equilibrium exists.

Section 4 introduces a variation to Banks and Sobel's Divine equilibria and presents an alternative iterative procedure to eliminate implausible sequential equilibrium based upon the interpretation of Bennett's theory of counterfactuals as it was introduced in the chapter I. There is a final section consisting in concluding remarks which analyzes the extent to which these refinements are plausible within the paradigm of rationality.

2. Signaling Games

Consider the following two players non-simultaneous move game: the first player, called the *sender* (S) chooses, after learning his *type* **t**, a *message* or *signal* **m**, from a finite set **M(t)**. This type is drawn from a finite set **T** according to a probability distribution π which is common knowledge among both players. The sender's type, that is the particular realization of π, is unknown to the second player, the *receiver* (R), whose decision consists in choosing an action, a, from a finite set **A(m)** after observing the sender's message, $m \in M(t)$. This response finishes the game. The resulting players' payoffs, **u(t,m,a)** and **v(t,m,a)** respectively, are determined by the message, the responding action and the sender's type.[37]

The rules of the game represented by $\Psi=(T,M,A,\pi,u,v)$ are common knowledge among the players. The asymmetry consists in the fact that the receiver, that

---

[37]As a simplifying assumption and without loss of generality we shall assume that M (t), the set of messages available to type t, is the same for all types. By the same token we shall also assume that A(m), the set of actions available after message m, is the same for all messages.

is player 2, does not know a piece of information that player 1 knows; namely player 1's type.

Let P(T), P(M) and P(A) be the set of probability distributions over T, M and A respectively. An element τ_m of P(T) represents a set of beliefs by player 2 concerning the likelihood of types $t \in T$ after receiving message m. Let $\underline{\tau}=(\tau_m)_m$ denote a system of beliefs of player 2. The elements of P(M) and P(A) will be denoted by μ and α respectively.

Let $\mathbf{p}=(\mathbf{p}_t)_t$ and $\mathbf{r}=(\mathbf{r}_m)_m$ respectively represent the sender's and receiver's behavioral strategies with $\mathbf{p}_t \in P(M)$ for all $t \in T$ and $\mathbf{r}_m \in P(A)$ for all $m \in M$. Given the sender's type t, p(. ;t) is a probability distribution over M(t) and given the sender's message, r(. ;m) is a probability distribution over A(m).

The expected payoff to type t when he sends the mixed message μ and player 2 responds with strategy r is

$$u(t,\mu,r) := \Sigma_{m,a}\ \mu(m)\, r_m(a)\, u(t,m,a).$$

On the other hand, the expected payoff of a player who faces message m when he has beliefs τ and responds with a mixed action α is

$$v(\tau,m,\alpha) := \Sigma_{t,a}\ \tau(t)\, \alpha(a)\, v(t,m,a).$$

Let p(m) be the probability that message m is chosen under strategy p:

$$p(m) := \Sigma_{t \in T(m)}\ \pi(t)\, p_t(m).$$

The beliefs of player who responds to message m calculated according to Bayes' rule are as follows:

$$\tau_m^{\,p}(t) := \pi(t) p_t(m) / p(m) \quad \text{if } p(m) > 0.$$

Let the best response by type t against r be,

$$BR_t(r) := \operatorname*{argmax}_{\mu}\ u(t,\mu,r),$$

and let the best response by player 2 to message m against beliefs τ be

$$BR_m(\tau) := \underset{\alpha}{\operatorname{argmax}}\ v(\tau, m, \alpha).$$

A Nash equilibrium for the game under consideration consists of a pair of behavioral strategies (p,r), such that

$$p_t \in BR_t(r) \quad \text{for all } t \in T,$$

and

$$r_m \in BR_m(\tau_m^p) \quad \text{for all } m \in M \text{ with } p(m) > 0.$$

In other words, under equilibrium the sender maximizes his expected utility given the receiver's response and, the receiver, after receiving the equilibrium message m, computes by Bayes' rule the probability that, given m, the sender is type t, for every t. In a second step, player 2 maximizes his expected utility using these probability assessments as his beliefs.

2.1 Sequential Equilibrium

There is general agreement upon the need of prescribing rational behavior at information sets that are off-the-equilibrium path for the equilibrium to be self-enforcing. The problem that remains, however, is how to model or think about these zero probability events. The concept of sequential equilibria is one of the notions that has been proposed to deal with this problem. Under this concept, players are modeled as expected utility maximizers *at all nodes*. Furthermore, in the face of uncertainty players are assumed to behave in the following way:

(i) they calculate conditional probabilities by Bayes' rule along the equilibrium path. This guarantees the consistency requirement upon beliefs;

(ii) they form beliefs off the equilibrium path by constructing posterior probability distributions on each information set that is not reached under equilibrium;

(iii) at each information set players assume that in the remainder of the game players will play according to the equilibrium under consideration.

A strategy profile (p,r) is a sequential equilibrium if there exist a system of consistent beliefs $\underline{\tau}=(\tau_m)_m$ such that each player's strategy maximizes his expected utility given his probability assessments on *and* off the equilibrium path.

Formally a triple $(p,r,\underline{\tau})$ is a sequential equilibrium if:

$p_t \in BR_t(r)$ for all $t \in T$,

$r_m \in BR_m(\tau_m)$ for all $m \in M$,

$\tau_m = \tau_m^p$ for all m with $p(m)>0$.

It is clear that a sequential equilibrium need not be optimal with respect to *all* beliefs. Moreover, the beliefs which support a sequential equilibrium need not be *sensible* A response to a deviation by the sender might be based upon the belief that the sender has played a dominated strategy. The reason is that deviations are implicitly treated as mistakes and therefore non connected with payoffs. Sequentiality *only* adds optimality at every node and consistency of beliefs along the equilibrium path.

2.2 The Intuitive Criterion

Consider the following game taken from Cho and Kreps [8]. Nature selects a type for player 1 who can be either *strong* (s) or *weak* (w). That is, T={s,w}. The probability distribution over the types is given by $\pi = \{\pi(s)=0.9, \pi(w)=0.1\}$. Without knowing whether player 1 is strong or weak and regardless of the message, player two has to decide whether to duel player 1 or not; that is A(m)=A={d, ~d}. If player 2

duels a strong player 1 he receives a payoff of 0; if he duels a weak player 1 he receives a payoff of 1. If player 2 decides not to duel his payoffs are 0 if the opponent is weak and 1 if he is strong. In other words player 2 would wish to duel if he believed his opponent was weak and not duel otherwise. After this response the game ends and the players receive their payoffs. In any event, player 1, who knows his type, has to choose between having beer or quiche for breakfast; that is, M(s)=M(w)={b,q}. Other things equal player 2's response, the weak type prefers a breakfast of quiche whereas the strong prefers a breakfast of beer. The payoffs to the players are depicted in Figure 1.

beer	duel	~duel	quiche	duel	~duel
strong	1,0	3,1	strong	0,0	2,1
weak	0,1	2,0	weak	1,1	3,0

Figure 2

This game has two set of Nash equilibria associated with the following outcomes:

1) player 1 regardless of his type has beer for breakfast; player two avoids the duel when the message is beer and otherwise duels with a probability of at least 0.5. In terms of our notation: p(b/s) = p(b/w) =1; r(~d/b)=1; r(d/q) ≥0.5 . This is indeed a sequential equilibrium provided that player 2's beliefs are such that τ_b^p (s)≥0.5 and τ_q (s) ≤ 0.5. The first inequality is guaranteed because along the equilibrium path, player 2's conditional beliefs are equal to the prior distribution (by Bayes' rule τ_b^p (s) = π(s)=0.9) and this prevents him from dueling. The second inequality is not constrained because there is no rule to compute these beliefs.

Existence of Nash equilibrium requires that player 2 duels if quiche with a probability of at least 0.5 because this response off the equilibrium path prevents the deviation by the first player regardless of his type. This Nash equilibrium is also sequential provided that player 2 believes that had player 1 had a breakfast of quiche he it would have been more likely that he was weak.

2) player 1 regardless of the type has quiche for breakfast; player two avoids the duel when the message is quiche and otherwise duels with a probability of at least 0.5. In terms of our notation: $p(q/s) = p(q/w) = 1$; $r(\sim d/q)=1$; $r(d/b) \geq 0.5$. This is indeed a sequential equilibrium provided that player 2's beliefs are such that $\tau_q^p(s) \geq 0.5$ and $\tau_b(s) \leq 0.5$. The first inequality is again guaranteed because by Bayes' rule $\tau_q^p(s) = \pi(s)=0.9$ and these beliefs prevent him from dueling. Off the equilibrium path, player 2 duels with a probability of at least 0.5 and this prevents the deviation by the first player regardless of his type. As before, this equilibrium is also sequential when player 2 believes that had player 1 had beer for breakfast it would have been more likely that he was weak.

Cho and Kreps argue that the equilibrium outcome in which both types drink quiche is not sensible. As we saw previously, this outcome is sequential because dueling if beer is a best response when the second player assigns a probability of at least 0.5 to the event that is the weak type the one who sent that message. Cho and Kreps argue that these beliefs are not sensible because the weak type can not profit from drinking beer relatively to the payoff which he receives under equilibrium if he has quiche. They assert that if this reasoning is common knowledge among the players, it is clear that player two should believe that the strong type is the more likely to have

deviated and therefore he should not duel. In this circumstance, the strong confirms the profitability of his deviation and this breaks the equilibrium outcome under analysis.

The formalization of this criterion can be described in the following way:

Let T(m) represent the set of types for whom message m is available. For each message m' sent off the equilibrium path define S(m') to be the set of types t whose equilibrium payoffs, u*(t), exceed the best payoff which they can possibly obtain if they deviate. In other words, $t \in S(m')$ iff:

$$u^*(t) > \max_{a \in BR(\tau(T(m')),m')} u(t,m',a)$$

An equilibrium outcome fails the *Intuitive Criterion* if there exist some type $t' \in T$ such that:

$$u^*(t') < \min_{a \in BR(\tau(T(m') \setminus S(m')),m')} u(t',m',a)$$

Roughly speaking, an equilibrium outcome fails the Intuitive Criterion when there is a type who, by deviating can profit relatively to the payoff that he obtains under equilibrium while facing a best reply to beliefs that exclude types who could never gain by deviating. In the beer-quiche game the strong type obtains a payoff of 2 under equilibrium. However he can reach a payoff of 3 if he deviates to beer and player two does not duel given his revised expectations that player 1 is the strong type.

However, it has been pointed out both by Cho and Kreps [8] and Van Damme [21] that if this reasoning is taken one step further and still assumed to be common knowledge among the players one concludes that *if* beer is a sure sign of a strong type *then* quiche is a sure sign of a weak type. By Bayes rule if $\tau(s/b)=1$ then $p(b/w)=0$. In other words, if beer signals the strong type then the weak type must have quiche for breakfast. Does this imply that quiche signals the weak type? To reply this question

with an affirmative answer we need to assume that the strong type does not have quiche for breakfast which is not equivalent to asserting that beer can only be chosen by a strong type. To imply that quiche signals the weak type we need to assume that the strong type apart from being rational believes that if he has beer the chance that he will face a duel is less than 0.5.

Cho and Kreps disclaim the argument that quiche signals the weak type by asserting that Nash equilibrium "is meant to be a candidate for a mode of self-enforcing behavior that is common knowledge among the players." They conclude that to test an equilibrium outcome one should start with the hypothesis that the corresponding equilibrium is common knowledge and then look for contradictions. In any case the conclusion that the weak type is better off by not having quiche leads to conclude that the quiche equilibrium outcome is not self-enforcing. However, one could also say that if the reasoning under consideration implies that beer does not signal a strong type then the initial hypothesis that beer signaled a strong type should be rejected instead of the equilibrium outcome itself. Van Damme suggests this type reasoning as a counter argument that beer off-the-equilibrium path signals the strong type. However he does not offer a way of solving this dilemma.

Cho and Kreps' opinion is valid if one thinks of an equilibrium as a recommendation to the players that guarantees them a certain payoff. Imagine that the players are told that if they have quiche for breakfast then it is certainly the case that the second player will not duel. Now they need to decide whether a deviation can improve the equilibrium payoff which they will receive with certainty. In this case Cho and Kreps' analysis results appropriate.

One can also think of an equilibrium as a set of consistent propositions within a language based upon a behavioral assumption and a theory of how to analyze deviations. Under this approach the logical consequences of the propositions that define the equilibrium should also be taken as part of it. From this point of view an equilibrium is a set of requirements whose consistency needs to be tested. An equilibrium can be disregarded as self-enforcing as long as we find an internal contradiction within this set of requirements. For instance the original quiche outcome is sequential and becomes inconsistent as soon as we consider the interpretation that beer signals a strong type. Without this further requirement that introduces constraints upon off the equilibrium path beliefs the quiche equilibrium outcome is internally consistent. Assuming that beer signals a strong type while fixing the equilibrium provides the strong type with incentives to deviate. However, if by assuming that beer signals the strong type, we accept as a logical consequence that quiche signals the weak type, then we give incentives to player 2 and the weak type to deviate from the original equilibrium. Either way, the internal consistency of the equilibrium outcome is lost.

One can also ask the following question: why should players have their equilibrium payoffs guaranteed and suppose that if they deviate *another* deviation will occur in response? or why should equilibrium responses to deviations be relaxed as an assumption and not equilibrium responses to equilibrium play? The framework developed in the first chapter provides the following answer. The factual world is that in which the equilibrium is played. Worlds in which deviations occur constitute counterfactual scenarios and one needs to drop at least one feature of the factual world in order to reach them. The problem is that within the framework of games there is no

unique way to depart from the equilibrium world in order to reach the counterfactual world of a deviation; many different departures might typically provide an access to circumstances in which a deviation takes place.

Finally it is worth noticing that when S(m)=T(m) for every message off the equilibrium path then there is no type that can be eliminated from the support of the beliefs of the second player. In this case the Intuitive Criterion results equivalent to the notion of sequentiality; that is, the criterion fails to refine the set of beliefs at nodes off-the-equilibrium path.

2.3 Divine equilibria

Consider the following game taken from Banks and Sobel [2]. Player one can be one of two types, called t_1 and t_2 with probabilities of 1/2 each. Each type has the same set of available messages; namely m_1 and m_2. The receiver has the same available actions after any of the messages: a_1 and a_2. The corresponding payoffs are depicted in Figure 3:

m_1	a_1	a_2	m_2	a_1	a_2
t_1	-3,3	-6,0	t_1	-5,5	-6,0
t_2	-3,3	-11,5	t_2	-5,5	-11,5

Figure 3

This game has two equilibrium outcomes:

$\{p(m_1/t_1) = p(m_1/t_2)=1$; r (a_1/m_1)=r $(a_1/m_2)=1\}$ and $\{p(m_2/t_1) = p(m_2/t_2)=1$; r $(a_1/m_2)=1$ r $(a_1/m_1)=0\}$. Both equilibrium outcomes survive the Intuitive Criterion.

However, Banks and Sobel claim that the equilibrium outcome that has both types sending message 2 is not sensible. The reason is that in order to support this equilibrium player 2 should believe that t_2 is more likely than t_1 and one can observe that whenever t_1 benefits from a response by player 2, t_2 benefits as well and not viceversa. We can describe this situation in the following manner: when m_1 is sent, the set of behavioral strategies by player 2 that outweighs the equilibrium payoff to t_1 contains the set of behavioral strategies by player 2 that outweighs the equilibrium payoff to t_2. Banks and Sobel assert that a sensible restriction to player 2's beliefs is that the probability of t_1 relative to that one of t_2 should increase when m_1 is received.

Their test can be summarized in the following manner. Let A_G be the subset of P(A(m')) such that type $t \in T$ obtains a payoff at least as good as the equilibrium payoff denoted by u*(t) when he sends message m'. Formally,

$$A_G(m') = \{\alpha \in P(A(m')): u(t,m',\alpha) \geq u^*(t) \text{ for some } t \in T\}$$

This is the set of actions by player 2 that type t prefers to equilibrium actions if he sends message m'. Banks and Sobel assume that the receiver should believe that the sender does not expect to lose from a deviation. Therefore the receiver should believe that type t expects him to take an action from A_G.

For all actions in P(A(m')) let λ be defined as:

$$\lambda(t,\alpha) = \begin{cases} 1 & \text{if } u(t,m',\alpha) > u^*(t) \\ [0,1] & \text{if } u(t,m',\alpha) = u^*(t) \\ 0 & \text{if } u(t,m',\alpha) < u^*(t) \end{cases}$$

$\lambda(t,\alpha)$ represents the probability that $t \in T$ would send m' if he believed that m' would induce action α assuming that he could have obtained the equilibrium payoff u*(t). Now let $\Gamma(m',\alpha)$ be the set of player 2's beliefs over the set of types T consistent

with player 2 taking action α in response to m' and type t obtaining u*(t) otherwise. Formally:

$\Gamma(m',\alpha) = \{\tau \in P(T): \exists\ \lambda(t) \in \lambda(t,\alpha)$ and $c>0$ such that $\tau(t)=c\lambda(t)\pi(t)\ \forall t \in T\}$.

Notice that this set is non empty if and only if $\alpha \in A_G$. Finally let

$\Gamma(A,m') = co[\cup_{\alpha \in A} \Gamma(m',\alpha)]$.

This set is empty only when $A_G \cap A$ is empty; this is when there is no type who can strictly benefit from a deviation considering all possible responses by player 2. When A_g is empty *any* conjecture supports the equilibrium. Otherwise Banks and Sobel assert that it is not plausible for player 2 to hold beliefs outside $\Gamma(A,m')$ given the signal m'. Conjectures in $\Gamma(A,m')$ assign zero probability to types who can never benefit from a deviation with respect to their equilibrium payoffs. Moreover, when $A_G(t,m') \subset A_G(t',m')$ for $t,t' \in T$ then for all beliefs in $\Gamma(A,m')$ the ratio of the probability of t' given m' to the probability of t given m' is at least as great as $\pi(t')/\pi(t)$.

Finally we can present the iterative procedure introduced by Banks and Sobel:

Let $\Gamma_0 = P(T)$, $A_0 = P(A)$ and for $n>0$

$\Gamma_n := \Gamma(A_{n-1})$ if $\Gamma(A_{n-1}) \neq \varnothing$ and $\Gamma_n := \Gamma_{n-1}$ otherwise.

$A_n := BR(\Gamma_n,m)$, $\quad \Gamma^* = \cap_n \Gamma_n$, and $\quad A^* = \cap_n A_n$.

A sequential equilibrium in a signaling game is *divine* if it is supported by beliefs in Γ^*. Returning to the game in Figure 2, the equilibrium outcome in which both types send m_2 is not divine[38]; in this case $\Gamma^* = \{\tau \in P(T): \tau(t_1)=\pi(t_1)=1/2\}$ and, as we already explained, these believes do not support player 2's off the equilibrium response in the case in which message m_1 sent.

---

[38] The only equilibrium outcome which is *divine* is the pooling equilibrium in which both types send message 1. In this case $A^* = a_1$.

This example also shows that the outcome of this procedure depends on the prior distribution π over the set of types T. With the purpose of overcoming this limitation Banks and Sobel redefine the set of beliefs that support the optimal response by player 2. Let Γ^{**} be the intersection of every Γ^{*} taken over all non degenerate priors on Sender types. A sequential equilibrium is *universally divine* if it is supported by beliefs in Γ^{**}. Naturally this is a more restrictive than divinity. In the game depicted in Figure 2 the pooling equilibrium in which message 2 is sent is divine only if $\pi(t_1) \leq 2/5$; however this equilibrium is not universally divine given that player 2 ought to believe that regardless of the prior the unexpected message comes from type 1.

2.4 The Intuitive Criterion and Divinity compared

Let us summarize the most important issues that regarding the methods presented in sections 2 and 3 have been discussed previously:

1) The interpretation of deviations as signals, which is an assumption in both the intuitive criterion and divinity, builds upon the idea that deviations can only emerge as a consequence of a rational decision. As it was asserted in the previous section this is one of the possible ways in which deviations can be interpreted. As it has been mentioned in the previous chapter, this approach to analyze off-the-equilibrium scenarios ought to be mutually shared by the players in order to ensure the convergence of their decisions and could lead to inconsistencies in some games.

2) In the beer-quiche game had a strong player 1 deviated from the quiche equilibrium he would have been better off in any of these two possible scenarios: *either* player 2 does not duel after beer *or* player 2 duels after quiche. The first case involves a deviation *off* the equilibrium path whereas the second a deviation *along* the equilibrium

path and therefore a rejection to the assumption that players are playing the equilibrium under analysis. On the other hand, a deviation by the weak player 1 could be profitable *only if* he expects player 2 to deviate from his equilibrium strategy after *every* possible message.

This means that the weak type can be justifiably eliminated only when we assume that player 2 responds to equilibrium play with equilibrium strategies. In other words, we need to fix the equilibrium under consideration and proceed *as if the reasoning mechanism*, which is common knowledge among the players, *had no further consequences upon the decision to play the equilibrium strategies*. Both the Intuitive Criterion and Divinity build upon this assumption. Moreover, once a player is eliminated, we consider player 2's best reply given the beliefs modified by this elimination. As it was already asserted, *if* under the quiche equilibrium beer signals a strong type *and* this implies that quiche signals a weak type *then* player 2 is better off by responding with a duel along the equilibrium path. This type of iteration is not possible under the methods described in sections 2 and 3 for the reasons given in the previous paragraph.

3) There are examples in which *every* type might potentially benefit from a deviation, even when the equilibrium is fixed. In this case the *Intuitive criterion* yields no further refinement upon the set of sequential equilibria. In some of these cases *Divinity* is capable of further restricting the set of sequential equilibria. To illustrate this consider the following example taken from Van Damme [21] :

m_1	a_1	a_2
t_1	2,2	2,2

m_2	a_1	a_2
t_1	-1,5	3,0

t_2 2,2 2,2 t_2 0,0 4,1

Figure 4

Banks and Sobel's *divinity* renders the equilibrium $\{p(m_2/t_1)=p(m_2/t_2)=1; r(a_2/m_2)=1\}$ as the only divine equilibrium provided that $\pi(t_2) >0.5$. The other equilibrium outcome is $\{p(m_1)/t_1)=p(m_1/t_2)=1; r(a_1/m_1)=1\}$which is not divine. In order to illustrate this result, fix the pooling equilibrium in which player 1 plays m_1. Banks and Sobel's iterative procedure starts by updating beliefs in the following way: consider the set of possible responses by player 2 which would make each type better off. In this case t_2 gains by deviating to m_2 only if $r(a_2/m_2)\geq 0.5$. On the other hand, t_1 benefits from a deviation to m_2 only if $r(a_2/m_2)\geq 0.25$. This implies that t_2 gains by deviating whenever t_1 does. Divinity requires that the belief that m_1 comes from t_2 be at least equal to the prior probability of this type. However this would make player 2 deviate to a_2 which eliminates this pooling equilibrium as a candidate for divinity.

Both equilibrium outcomes of this game pass the test or speech proposed by Cho and Kreps' whereas only $\{p_1(m_2)= p_2 (m_2)=1; r_{m2} (a_2)=1\}$ is divine.

Finally, it is worth noticing that *Divinity* depends upon the prior distribution over types. In the game depicted in figure 4 there is only one divine equilibrium provided that $\pi(t_2) >0.5$. However, when $\pi(t_2) \leq 0.5$ both equilibrium outcomes are divine.

3 Priors as off-the-equilibrium beliefs

Equilibria in signaling games of the sort described in section one can be typically of two different types: either pooling or separating. In the first case each type

of player 1 chooses the same message and this implies that player two does not learn anything about the type of his opponent along the equilibrium path. This is due to the fact that under this circumstance Bayes' rule renders the conditional posterior of the types equal to the prior distribution. In the second case, messages perfectly signal player 1's type and in this way player 2 knows with certainty which type of player 1 he is facing.

Consider again the beer-quiche game. In order to support the equilibrium in which both types have beer for breakfast it is necessary that if player 1 has quiche for breakfast player 2 duels with a probability of at least 0.5. In order to do so, player 2 needs to have beliefs that assign a probability of at least 0.5 to the weak type conditioned on the observation of quiche. In order to avoid the duel *along* the equilibrium path, player 2 needs beliefs that assign a probability of at least 0.5 to the event that he faces a strong player 1 conditioned on the observation of beer. Given that under this equilibrium $\tau_b^p(s):=\pi(s) = 0.9$, player 2 has no incentives to deviate along the equilibrium path.

However, if player 2 uses the prior distribution over types as his beliefs in case of a deviation he should not duel either and this clearly breaks the equilibrium under consideration.

Using this criterion and given the structure of this game, existence of Nash equilibrium would be guaranteed only if $\pi(s)=\pi(w)=0.5$. That is, existence is guaranteed only when the priors equal the threshold probability that makes player 2's indifferent between the two different responses along and off the equilibrium path.

The equilibrium in which both types have beer for breakfast is supported only when quiche signals that a weak type is more likely than a strong type. When player 2

does not update his beliefs in this way, the weak type will have clear incentives to deviate to a breakfast of quiche. The pooling equilibria of this game requires that the two sets of player 2's beliefs that respectively support his response on and off the equilibrium path have only one element in common, which we called the threshold. The crucial feature that guarantees this is that neither player 2's payoffs nor his available actions depend on player 1's message. It is clear that in this situation and under the assumption that on and off the equilibrium path beliefs are equal to the prior distribution over the types, Nash equilibrium exists only when this prior distribution equals the threshold beliefs over types that makes player 2 indifferent between his available actions.

Consider the pooling equilibrium outcome in the game depicted in Figure 3 in which both types send message 1 and player 2 replies with a_1. Regardless of player 2's beliefs, a_1 is always a best response (at least as good as a_2) if he faces message 2. Therefore in this case any prior used as beliefs off the equilibrium path will support the equilibrium. Consider now the pooling equilibrium in which both types send m_2. If player 2 responds with a_2 off the equilibrium path then no type will wish to deviate. This implies that player 2 should have beliefs that attach a probability of at least 3/5 to type t_2. If player 2 uses the prior distribution as his beliefs, any prior distribution such that $\pi(t_2)\geq 3/5$ supports the equilibrium outcome under consideration.

As it was argued in Chapter I deviations represent counterfactual scenarios that can not be uniquely interpreted unless a theory of how to attach meaning to deviations is introduced. Two alternatives were proposed to model deviations motivated by two different theories of counterfactuals. Under the first interpretation, motivated by Lewis's theory of counterfactuals, deviations were modeled as thought experiments in

the sense of being options available to the players that they can scrutinize even though under equilibrium they constitute irrational choices. In other words, deviations were not supposed to confer a signal. The crucial question is how should player 2 update his beliefs concerning the type and rationality of his opponent if deviations are meaningless. A sensible solution is to have player 2 adopting the piece of information that seems to be more reliable. Namely the prior distribution of types.

3.1 The role of payoffs and priors in the existence of equilibrium

The drawback of the methodology outlined in the previous subsection is that existence of equilibrium can not be guaranteed. The restriction imposed upon the set of beliefs over the types in case of a deviation might be too strong given the payoff structure of the game to allow for an equilibrium. When the prior distribution over the types is taken as the beliefs of the second player in a signaling game, existence of a pooling equilibrium is only guaranteed for a strict subset of priors if we fix the payoff structure. However, given that players decide upon their play by considering their expected utility it should also be noted that the payoff structure albeit typically fixed also plays a role. To analyze this interaction let us consider the basic case of two types of player 1, two messages and two replies by player 2:

m_1	a_1	a_2
t_1	u^1_{11},v^1_{11}	u^1_{12},v^1_{12}
t_2	u^1_{21},v^1_{21}	u^1_{22},v^1_{22}

m_2	a_1	a_2
t_1	u^2_{11},v^2_{11}	u^2_{12},v^2_{12}
t_2	u^2_{21},v^2_{21}	u^2_{22},v^2_{22}

Figure 5

Without loss of generality assume that the pooling Nash equilibrium of this game is: E ={m_1, m_1, a_2, a_1}. Player 2 prefers a_2 to a_1 after receiving message m_1 if and only if:

$$v^1_{11}\,\tau_{m1}(t_1) + v^1_{21}\,[1-\tau_{m1}(t_1)] \le v^1_{12}\,\tau_{m1}(t_1) + v^1_{22}\,[1-\tau_{m1}(t_1)] \qquad (3.1.1)$$

Along the equilibrium path $\tau_{m1}(t_1)=\pi(t_1)$. Therefore we can rewrite (3.1.1) as:

$$v^1_{11}\,\pi(t_1) + v^1_{21}\,[1-\pi(t_1)] \le v^1_{12}\,\pi(t_1) + v^1_{22}\,[1-\pi(t_1)] \qquad (3.1.2)$$

On the other hand, player 2 prefers a_1 to a_2 after receiving message m_2 if and only if:

$$v^2_{11}\,\tau_{m2}(t_1) + v^2_{21}\,[1-\tau_{m2}(t_1)] \le v^2_{12}\,\tau_{m2}(t_1) + v^2_{22}\,[1-\tau_{m2}(t_1)] \qquad (3.1.3)$$

Assuming that $\tau_{m2}(t_1)=\pi(t_1)$ we rewrite (3.1.3) as:

$$v^2_{11}\,\pi(t_1) + v^2_{21}\,[1-\pi(t_1)] \le v^2_{12}\,\pi(t_1) + v^2_{22}\,[1-\pi(t_1)] \qquad (3.1.4)$$

Equations (3.1.2) and (3.1.4) provide two constraints for the values or $\pi(t_1)$ such that an equilibrium exists. It is clear that equilibrium exists if only if the actual value of $\pi(t_1)$ satisfies both equations and this at least requires that the intersection of these two bounds be non empty.

Although the payoffs in equations (3.1.2) and (3.1.4) do not overlap the assumption that posterior beliefs equal prior probabilities introduces a link between the responses to different messages received by the second player. In other words,

although replies to every message are based upon different payoffs they should be motivated by compatible beliefs.

In the beer quiche game (Fig 2) equations (3.1.2) and (3.1.4) provide the following bounds for the equilibrium in which both types drink beer: $\pi(s) \leq 0.5$ and $\pi(w) \geq 0.5$ respectively. Therefore equilibrium would exist if and only if $\pi(t_1)=0.5$. In the game depicted in Figure 3, if one considers the pooling equilibrium in which both types send m_2, equations (3.1.2) and (3.1.4) respectively require that $\pi(t_1) \geq 0$ and $\pi(t_1) \leq 0.4$. Thus equilibrium exists for $0 \leq \pi(t_1) \leq 0.4$.

4 A variation of Banks and Sobel's *Divinity*

Consider the game in Figure 6 with $\pi(t_1)=0.9$ $\pi(t_2)=0.1$.

	m_1	a_1	a_2		m_2	a_1	a_2
	t_1	0,0	0,0		t_1	-1,0	1,1
	t_2	0,0	0,0		t_2	-1,1	1,0

Figure 6

This game has two equilibrium outcomes: $\{p_1(m_1)= p_2(m_1) =1; r_{m2}(a_1)=1\}$ and $\{p_1(m_2)= p_2(m_2) =1; r_{m2}(a_2)=1\}$. Consider now the pooling equilibrium where both types send message m_1. No type can be eliminated by the *Intuitive Criterion*. Neither does *divinity* refine the set of sequential equilibria because all of them are divine. The reason is that both types can potentially benefit from a deviation in exactly the same circumstances; that is, $A_G(t_1,m')=A_G(t_2,m')$. Therefore $\Gamma^*=P(T)$.

Consider now the following variation to the iterative procedure that defines divine equilibria:

$\underline{\lambda}(t,\alpha)= 1$ if $u(t,m',\alpha) \geq u^*(t)$ and

$\underline{\lambda}(t,\alpha)= 0$ otherwise.

$\underline{\Gamma}(m',\alpha) = \{\tau \in P(T);$ such that $\tau(t)=c\lambda(t)\pi(t)\ \forall t\in T$ and $c>0\}$.

Moreover when $A_G(t,m')$ is empty for all t in T and therefore $\underline{\Gamma}(A,m')=\varnothing$ assume that

$\underline{\Gamma}(m',\alpha) = \{\tau \in P(T);$ such that $\tau(t)=\pi(t)\ \forall t\in T\}$.

Calculate the remaining by following the iterative procedure outlined in Section 2.3 replacing Γ and λ by $\underline{\Gamma}$ and $\underline{\lambda}$ respectively.

The pooling equilibrium of the game depicted in Figure 6 in which both types play m_1 is supported by this variation of the procedure provided that $\pi(t_1)\leq 0.5$. With the modification presented above player 2 uses the priors as his beliefs when he observes the off-the-equilibrium message m_2 and therefore decides to play a_1. This response deters both types from deviating.

4.1 The refinement to Divinity and Divinity compared

In this section we compare the variation to Divinity just outlined with Banks and Sobel's iterative procedure in order to characterize the former concept of equilibrium.

Let us consider again a signaling game in which there are two types of player 1, two available messages for each type and two available responses by player 2. Without loss of generality consider the following pooling equilibrium: $E =\{m_1, m_1, a_2, a_1\}$. In

this equilibrium both types send m_1; player 2 responds with a_2 to this message and with a_1 to m_2.

Given the payoffs to the first player, four possible scenarios are feasible regarding the set of actions by player 2 that each type prefers to equilibrium actions:

i) $A_G(t_1)=A_G(t_2)=\varnothing$.

In this case no type can potentially gain by deviating relative to his equilibrium payoff and this implies that $\Gamma(A_{n-1})=\varnothing$ for all $n>0$. Moreover $\Gamma_n=P(T)$ for all $n>0$ and $\Gamma^*=\cap_n\Gamma_n=P(T)$. This means that every sequential equilibrium is Divine.

Regarding the variation presented in the previous subsection, $\underline{\Gamma}(m',\alpha)=\{\tau(t_1)=\pi_1; \tau(t_2)=\pi_2\}$. This implies that $\underline{\Gamma}^*$ is a singleton consisting of the prior distribution over the types. An equilibrium satisfies the test presented in section 4 if and only if the prior distribution over types satisfies the boundaries defined in (3.1.2) and (3.1.4).

ii) $A_G(t_1)=\varnothing$; $A_G(t_2)\neq\varnothing$ (alternatively $A_G(t_2)=\varnothing$; $A_G(t_1)\neq\varnothing$).

In this case there is only one type who might benefit from a deviation and this leads player 2 to believe that the deviation certainly comes from this type: $\Gamma^*=\{\tau(t_1)=0; \tau(t_2)=1\}$(alternatively $\Gamma^*=\{\tau(t_1)=1; \tau(t_2)=0\}$). In this particular case, the refinement that Divinity imposes over sequentiality is equivalent to that imposed by the Intuitive Criterion. The beer-quiche game depicted in Figure 2 illustrates this case.

The variation of Divinity outlined in the previous subsection is equivalent in this case to Divinity; that is, $\underline{\Gamma}^*=\Gamma^*$.

iii)$A_G(t_1)=A_G(t_2)\neq\varnothing$.

In this circumstance the set of actions by player 2 that both types prefer to equilibrium actions coincides. This means that both types could potentially benefit from a deviation in the exact same situations. As in case i) $\Gamma^*=\cap_n\Gamma_n=P(T)$ and therefore

Divinity is equivalent to sequentiality: an equilibrium is Divine if and only if it is sequential. In the game introduced in Figure 6 the set of equilibrium outcomes such that both types send m_1 is an example of this case.

The modified set of beliefs $\underline{\Gamma}^*$ by the second player is a singleton in this situation consisting of the prior distribution over the types as in case i). As before, an equilibrium satisfies the test presented in section 4 if and only if the prior distribution over types satisfies the boundaries defined in (3.1.2) and (3.1.4).

iv)$A_G(t_1)\neq A_G(t_2)$; $A_G(t_1)\neq\varnothing$; $A_G(t_2)\neq\varnothing$.

Although both types might potentially gain relative to their equilibrium payoffs, there are responses by player 2 that would induce a deviation by only one of the types. That is, either $A_G(t_1)\subset A_G(t_2)$ or $A_G(t_2)\subset A_G(t_1)$. The game depicted in Figure 3 illustrates the case in which $A_G(t_2)\subset A_G(t_1)$. As we already saw not every sequential equilibrium is Divine in this circumstance although every sequential equilibrium satisfies the Intuitive Criterion.

The variation of Divinity outlined in section 4 is equivalent in this case to Divinity; that is, $\underline{\Gamma}^*=\Gamma^*$.

5. Concluding remarks

The goal of this chapter has been to analyze different ways of updating off the equilibrium path beliefs in relation with the concept of sequential equilibrium in signaling games.

It follows from the analysis developed in the first chapter that there is no unique way to model off the equilibrium behavior and that certain approaches are not always compatible with the basic assumption of common knowledge of rationality. There are

different ways through which the counterfactual world of a deviation might be reached and the different alternatives trade off with each other as a possible explanation. All these alternativess involve the relaxation of at least one of the assumptions that hold in the equilibrium world where players do not deviate. There are on the one hand the assumptions concerning the rationality of the players (including the possibility of their making a mistake) and on the other, the assumptions concerning the amount of knowledge that players possess about the structure of the game and the rationality of their opponents.

This is an important issue because the way in which beliefs are updated after a deviation together with the resulting equilibrium outcome crucially depend on the explanation that undelies the deviation. A possible solution is to link the interpretation of deviations to the structure of the game when rationality is the last assumption that the theorist might want to relax. When deviations might not conceivably lead to a potential gain, intentionality can not be compatible with rationality. In this case deviations might be considered meaningless. In this situation it seems reasonable that players base their responses upon the information which is common knowledge in the game and proceed as if no further deviations were expected. Alternatively one could also assume that at least one of the pieces of information concerning the structure of the game is not common knowledge. For instance it can be supposed that players are guided by different payoffs from the ones their opponents expect them to have. On the other hand, when the payoff structure is such that some player might potentially benefit from not conforming to his equilibrium strategy, it may be assumed that his deviation signals either the player's future play or reveals some information which was not common knowledge before the deviation.

Signaling games constitute a good example to apply these criteria. We have proposed that signals exist only when they are not incompatible with the assumption of common knowledge of rationality. Messages by a player who is more informed than his opponent might signal the unknown piece of information in the face of a deviation only when the corresponding deviation can be rationally explained. The refinements considered in this chapter, namely The *Intuitive Criterion* and *Divinity* refine the set of sequential equilibrium in some of these circumstances. The first refinement is effective when there are players whose deviations can not be rationally explained. The second refinement is effective not only in this circumstance but also when the players who might deviate can be separated in terms of the responses that they would prefer after a deviation. We have propose a variation of this second refinement aimed at determining the beliefs of the uninformed player when the different types who send messages can not be discriminated in terms of their propensity to deviate.

We have also proposed that when deviations are not compatible with common knowledge of rationality they be considered meaningless and therefore assumed to imply no further updating in the beliefs of the player who responds to the deviation. A case has been made in the present chapter for the use of prior probabilites as the ex post beliefs after a deviation. This is justified by the fact that the prior distribution over types is a piece of information which is common knowledge ex ante and guides players' responses along the equilibrium path. The drawback is that in this situation equilibrium might not exist as it has been shown. This illustrates the well known trade off between existence of equilibrium and the extent to which it can be refined.

III. References

[1] Aumann, Robert "Backwards Induction and Common Knowledge of Rationality", Games and Economic Behavior 8, 6-19 (1995).

[2] Banks, J. & Sobel, J. "Equilibrium Selection in Signaling Games" Discussion Paper # 85-9, mimeo, Department of Economics University of California, San Diego.

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

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

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

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

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

[8] Cho,I.K. & Kreps,D. "Signaling Games and Stable Equilibria" The Quarterly Journal of Economics, vol.CII; May 1987 Issue 2.

[9] Elster, Jon Rational Choice New York university Press, 1986 pp. 13.

[10] Fudenberg, Drew and Tirole, Jean Game Theory MIT Press, 1991.

[11] Goodman, Nelson Fact Fiction and Forecast Harvard University Press 1979,1983.

[12] Harper, William "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.

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

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

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

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

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

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

[19] Selten, Reinhard and Leopold, Ulrike "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, Robert "A Theory of Conditionals" in IFS edited by Harper, W., Stalnaker,R. and Pearce, G. D.Reidel Publishing Company.

[21] Van Damme, Eric "Stable equilibria and forward induction" Journal of Economic Theory 48, 476-496.

[22] Van Damme, Eric "Stability and Perfection of Nash Equilibria" Springer-Verlag 1987,1991.