Backwards Forward Induction

Gian Aldo Antonelli
Yale University

Cristina Bicchieri
Carnegie Mellon University

December 1993

Abstract

In this paper we isolate a particular refinement of the notion of Nash equilibrium that is characterized by (*i*) the fact that it provides a unified framework for both backwards and forward induction; and (*ii*) by the fact that it is mechanically computable. We provide an effective procedure, whose definition embodies certain given principles of rationality. Such a procedure allows us, given a representation of an extensve form game, to compute a set of paths through the tree corresponding to what we call "reasonable equilibria." The reasonable equilibria are all shown to be Nash equilibria. Further, such a notion of reasonable equilibrium agrees with backwards induction in the case of extensive form games of perfect information, and with forward induction in the case of extensive form games of imperfect information. This allows us to model the players' reasoning process by giving a theory (with which each player is supposed to be endowed), from which statements characterizing the players' behavior are deducible. Such a theory is not yet *complete*, in that it cannot handle true (irrational) deviations. We point at directions for future work by showing how such a theory can be made complete provided we re-interpret some of its axioms as defeasible inference rules.

1 Introduction

Two problems have been widely discussed in recent debates on the foundations of game theory, namely: (i) the problem of how to handle deviations from equilibrium play, and (ii) the problem of how to model counterfactual reasoning patterns that appear to be necessary to carry out backwards induction (Aumann [1], Stalnaker [12]). There are obvious connections between these two problems, although they have usually been treated separately.

The first problem appears in the context of games that have multiple Nash equilibria, some of which might be implausible in that they involve *risky* (i.e. weakly dominated) strategies and the implausible beliefs that those strategies will be played. Various refinements of Nash equilibrium have been proposed to take care of implausible equilibria, as well as to attain predictability in the face of multiplicity. In games of perfect information this is accomplished by applying backwards induction. In games of imperfect information instead, one has to appeal to different types of refinements, each of which corresponds to a different way of checking the stability of a Nash equilibrium against deviations from equilibrium play. The stability of an equilibrium, however, is a function of how a deviation is being interpreted. Counterfactuals play a role in this context since, from the viewpoint of a particular equilibrium, an off-equilibrium move is a contrary-to-fact event. When considering the possibility of such an event, a player has to undergo a belief-revision process, retracting from his original set of beliefs all those beliefs that contradict the statement that an off-equilibrium event has taken place. The interpretation of out-of-equilibrium play will thus depend on the model of belief revision adopted (and on the interpretation of the counterfactual), and so will the resulting refinement (Bicchieri [2, 3, 4]).

As a refinement for games of perfect information, backwards induction embodies the principle that a rational player will only play undominated strategies. In this context, too, the issue of counterfactuals arises when a player contemplates a deviation from equilibrium play. But whereas in games of imperfect information a deviation may or may not be inconsistent with common knowledge of rationality, in a large class of games of perfect information deviations are inconsistent with rationality being common knowledge (Bicchieri & Antonelli [5]). Augmenting the theory of the game with an account of counterfactuals can solve the problem, at a price. Different kinds of game may need different accounts of counterfactuals to make deviations compatible with rational behavior being common knowledge, whereas one would like to have a unique general account of counterfactuals that applies to all types of games.

It is often argued that a complete theory of the game is a theory that explains the unexpected, that is, it is a theory that explains all sorts of moves, including irrational ones, on the part of the players. The task of constructing such theories, however, has proved quite formidable. For example, a theory that interprets deviations as mistakes takes into account all sorts of deviations, but, as we shall show in section 2, a theory of mistakes may be incompatible with rationality being common knowledge. On the other hand, a theory that only considers rational deviations, interpreting them as signals, is consistent with rationality being common knowledge, but it is silent as far as irrational deviations are concerned. In the present paper, we shall argue that the treatment of deviations, be they studied in finite extensive form games of perfect or imperfect information, does not require a full account of (intra-theoretical) counterfactual reasoning, notoriously one of the thorniest issues in philosophical logic and knowledge representation.

In our model, the players reason to an equilibrium on the basis of the theory of the game they are endowed with. Players are like automatic theorem provers provided with a decision procedure that isolates, whenever possible, a unique equilibrium that satisfies the rationality conditions embedded in the theory's axioms. A surprising result of such a decision procedure is that it leads quite naturally to the backwards induction equilibrium in finite games of perfect information, and to the forward induction refinement in finite games of imperfect information.

The theory of the game T is formalized in classical first-order logic and is a revised version of Bicchieri & Antonelli [5]. Our theory, which is interpretable in Primitive Recursive Arithmetic, comprises general axioms describing the game (represented by a finite tree) and the payoff structure for each player. We supply a function π^* giving, for each information set, the set of undominated paths starting at that set. Finally, we give "behavioral" axioms describing how players' actions are determined by the set of undominated paths and hence, indirectly, by their expected payoff at each information set. This theory allows players to infer the sequence of moves comprising what we call a Reasonable Equilibrium path (or paths), i.e. an equilibrium path that satisfies the rationality conditions that are embedded in our theory.

The theory we propose here differs from Bicchieri & Antonelli [5] in several important respects. It is far more general, as it also includes games of imperfect information, and it is not assumed that players have "local knowledge" at an information set; in fact, it is assumed that the theory T is group knowledge among the players[1]. Since we model a decision procedure that leads to the isolation of a subset of undominated paths, no belief revision is needed to identify the subset of Nash equilibria that contains only undominated strategies. Belief revision is only necessary when a player deviates to a dominated path. In our view, it makes sense to call "deviation" only an action that is unexpected in that it is obviously contrary to the best interest

[1] By group knowledge of p we mean that every member of the group knows p

of the deviant player. Such action should then be explained.

The theory T we assign to the players only provides a background description of the game. Considering a deviation means augmenting T by a "history," i.e. by axioms $A_1, \ldots, A_k$ specifying that certain moves have been made. Since those moves lie outside the undominated path(s), $T + A_1 + \ldots + A_k$ is inconsistent. This implies that any model of belief revision based on T and meant to accommodate deviations cannot but generate a modification of T.

By exhibiting theory T we accomplish a twofold task: First, we show how a first-order theory of the game is perfectly adequate to infer the backwards induction equilibrium in games of perfect information and, in games of imperfect information, it gives us a refinement that agrees with forward induction. These results are obtained without an account of counterfactuals. Second, we set the stage for a default version of the theory obtained by the original theory T by weakening the axioms, i.e. by reinterpreting certain material implications as defeasible default rule. A default theory of the game is a complete theory in the sense that it can be augmented with information to the effect that a *true* deviation (i.e. an irrational move) from an equilibrium path has occurred without becoming inconsistent.

2 An Example

In the refinements literature, anticipated actions off the equilibrium path play a crucial role in sustaining an equilibrium. Given a Nash equilibrium, the players are supposed to ask what would happen if one of them were to deviate from the equilibrium path, and an equilibrium is considered plausible (or stable) only in case the players would have no incentive to play another strategy in face of a deviation. Form the viewpoint of a given Nash equilibrium, asking what to do when a deviation occurs is tantamount to asking a counterfactual question, since an off- equilibrium move is by definition a contrary-to-fact event. Though most game theorists now recognize that the treatment of deviations involves counterfactual reasoning and a change of beliefs on the part of the players, there are very few syntactical or semantical models of belief change in the literature. This is mainly due to the fact that the model of the game being played is the game theorist's and not the players'. This also explains why so little attention has been paid to whether the beliefs attributed to the players are reasonable, in the sense of being consistent with their information about the game. What follows is

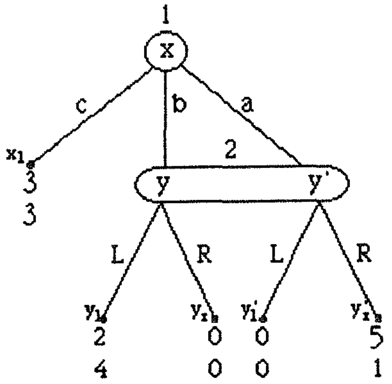

Figure 1: A two-person game of imperfect information.

an example of how the problem of multiple equilibria is usually addressed, where the kinds of refinement proposed depend upon the beliefs attributed to the players.

The game in Figure 1 is a two-person game of imperfect information. As usual, the players are assumed to have common knowledge of rationality (i.e. that they are expected utility maximizers) and of the structure of the game. Player 1 has three choices: either he chooses c, which ends the game, or he may choose b or a, in which case it is player 2's turn to move. If player 2 is called upon to move, however, she will not know player 1's preceding move, so she cannot tell whether she is at node y or y'. The game has two Nash equilibria in pure strategies, (c, L) and (a, R), and one would like to find some means to predict which equilibrium will be chosen by the players. Yet in such a simple game the most common refinement concepts, such as perfection, properness, or sequential equilibrium do not succeed in selecting a unique equilibrium. Let us see why.

The equilibria (c, L) and (a, R) are both perfect (Selten [11]). In particular, (c, L) is perfect if player 2 believes that 1 will make mistake b with a greater probability than mistake a, but whereas both probabilities are very small, the probability of playing c is close to 1. If this is player 2's belief, then she should play L with probability close to 1. In this case, player 1 should play c. But why should 2 believe that mistake b is more likely than mistake a? Since strategy c strictly dominates b, there is no reason to expect mistake b to occur more frequently than mistake a. The beliefs that support (c, L) are thus unreasonable.[2] The problem is that out-of-equilibrium beliefs are unrestricted: A player is supposed to ask whether it is reasonable to believe that the opponent will play his part in a given Nash equilibrium, but not whether the beliefs supporting the opponent's choice are rational. A rational belief, in this case, means a belief consistent with rationality being common knowledge. In our example, player 1 attributes to player 2 a belief that justifies her choice of strategy L, but it is not obvious that 2's belief about the greater likelyhood of mistake b is defensible.

It could be argued that one way to restrict out-of-equilibrium beliefs is to restrict a player's conjectures about the opponent's behavior to those that are rationally justified. A rational player, for example, should be expected to avoid costly mistakes (Myerson [8]). Proper equilibria need only be robust with respect to plausible deviations, i.e., deviations that do not involve costly mistakes. However, one mistake may be more costly than another only insofar as the player who could make the mistake has definite beliefs about the opponent's reaction. In our example, both (c, L) and (a, R) are proper equilibria. If a deviation from (c, L) were to occur, player 2 would keep playing L only if she were to assign a higher probability to mistake b than to mistake a. And if player 1 were to expect 2 to play L, mistake b would indeed be less costly than a. In this case, L would be a best reply for player 2. Thus mistake b is less costly if 1 expects 2 to play L, and 2 will play L only if she believes that 1 expects her to play L in response to a deviation. But, again, why should player 2 be expected to play L in the first place? Since b is strictly dominated by c, it is very unlikely that b occurs. The only plausible deviation is thus a, but then player 2 should be expected to play R.

The same problem arises with the sequential equilibrium notion (Kreps & Wilson [7]), which explicitly specifies beliefs at information sets lying off the equilibrium path. In our example, both (c, L) and (a, R) are sequential equilibria, since an equilibrium strategy has to be optimal with respect to some beliefs, but not necessarily plausible beliefs. In particular, if player 1 chooses c, then any probability assessment by player 2 is reasonable, and it is entirely possible that player 2 assesses a higher probability to strategy b than to a.

The equilibrium (c, L) is intuitively unreasonable precisely because the beliefs that support it are unreasonable. By reasonable beliefs we mean

[2] Even Selten [11, p. 35] admits that game theory is concerned with absolutely rational agents and that "there cannot be any mistakes if the players are absolutely rational."

beliefs that are consistent with common knowledge of rationality. If rationality is common knowledge, a player should never be expected to choose a dominated strategy. This must be true of weakly dominated strategies, too. The rationale for this condition is simple: Since off-equilibrium choices are relevant only when they affect the choices along the equilibrium path, it is reasonable to ask that an off-equilibrium choice that is weakly dominated should be ruled out, since it is as good as some other strategy if the opponent sticks to the equilibrium, but it does worse when a deviation occurs. In our example, rationality is common knowledge and strategy b is strictly dominated, therefore it must be common knowledge that b is never going to be played. Knowing that player 2 will always respond to a deviation with R, player 1 will have an incentive to choose a. This kind of reasoning rules out (c, L) as implausible.

Considering only undominated choices means that off-equilibrium beliefs should satisfy the following condition:

(R) When considering a deviation from a given equilibrium, a player should not hold beliefs that are inconsistent with common knowledge of rationality.

All that condition **(R)** tells us is that whenever a player has a weakly dominated strategy he should not be expected to use it, and that no one should choose a strategy that is a best reply to a weakly dominated strategy. Common knowledge of rationality thus implies common knowledge that weakly dominated strategies will not be used. Note that condition **(R)** entails iterated elimination of dominated strategies in the strategic form. Consider the strategic form of the game in Figure 1, which is given in Figure 2. In this game b is eliminated since it is strictly dominated by c. Since b is eliminated, R weakly dominates L, which is in turn eliminated. Finally, a dominates c for player 1, hence (a, R) is the only equilibrium that survives iterated elimination of dominated strategies.

Is there a correspondence between the iterated procedure we have just described and our informal argument in favor of (a, R) in the extensive form? If the game in Figure 1 were one of perfect information, backwards induction would give us a decision procedure that matches iterated elimination in the strategic form. Starting from terminal nodes, players eliminate weakly dominated strategies bottom up; in the absence of ties, this method determines a single outcome. In our example, it would be the equilibrium (a, R). Note that backwards induction requires rational behavior even in those parts of

	L	R
c	3,3	3,3
b	2,4	0,0
a	0,0	5,1

Figure 2: The strategic form of the game of Figure 1.

the tree that may not be reached if an equilibrium is played. As a result, backwards induction leads to eliminate all but the equilibrium points that are in equilibrium in each of the subgames and in the entire game. More generally, we may state the following backwards induction condition:

(BI) A strategy is optimal only if that strategy is optimal when the play begins at any information set that is not the root of the game tree.

In games of perfect information, **(R)** and **(BI)** guarantee that unreasonable equilibria are ruled out. Together, they imply that a plausible Nash equilibrium must be consistent with deductions based on the opponent's rational behavior in the future. Future behavior, however, may involve off-equilibrium play, and in this case condition **(R)** tells us that the only deviations that matter are undominated choices, i.e., choices that can be interpreted as intentional moves of rational players.

The game in Figure 1, however, is one of imperfect information; here the backwards induction algorithm fails because it presumes that an optimal choice exists at every information set, given a specification of play at the successors. At 2's information set, however, there is no unique rational action: at node y she should play L, and at node y' she should play R. Even if backwards induction is not defined, conditions **(R)** and **(BI)** may still apply to games of imperfect information that have proper subgames. For each such subgame, one may ask whether an equilibrium for the whole game induces an equilibrium in the subgame (Selten [10]). Yet the game in Figure 1 has no proper subgames, so in this case **(BI)** does not apply. Condition

(R) still applies, though, by constraining the possible interpretations of deviations.

In Figure 1, if player 2 gets to play then player 1 must have foregone the payoff of 3 in favor of playing a. The only equilibrium that yields a payoff greater than 3 to player 1 is (a, R), hence 2 should deduce from the fact that her information set is reached that 1 has chosen strategy a. If so, 2's best reply is R and player 1, anticipating player 2's reasoning, will conclude that it is optimal for him to play a. What we have just described is a forward induction argument, that is, an argument based on inferences about the opponent's rational behavior in the past. In our example there is no past to speak of, but rather the knowledge that player 1, facing the choice of getting a payoff of 3 for sure or playing a simultaneous game with player 2, has chosen the second option. A forward induction argument thus interprets deviations from a given equilibrium as signals, intentional choices of a rational player (Kohlberg & Mertens [6]). For this interpretation of deviations to be consistent with rationality, however, there must exist at least a strategy that yields the deviating player a payoff greater or equal to that obtained by playing the equilibrium strategy. This consideration leads to the following iterated dominance requirement:

(ID) A plausible equilibrium of a game G must remain a plausible equilibrium of any game G' obtained from G by deleting (weakly) dominated strategies.

Condition **(ID)** implies the iterated use of condition **(R)** in games that have subgames. Taken together, conditions **(ID)**, **(R)** and **(BI)** underlie the forward induction argument. Consider the following game: In Figure 3 each player has the choice of playing down, which ends the game, or playing across. At node w, if player 1 chooses to play across he plays a simultaneous battle of the sexes with player 2. This game has two equilibria in pure strategies: (A_1A_3T, D_2) and (A_1D_3, A_2R). Note that in the strategic form of Figure 3 the equilibrium (A_1D_3, A_2R) does not survive iterated elimination of (weakly) dominated strategies.

Like other refinements, forward induction is used to check the equilibria of the game against possible deviations. The difference with other refinements lies in the criteria used to assess deviations. Suppose the players agree to play (A_1D_3, A_2R) but, unexpectedly, player 1 deviates at node w by playing A_3. Since A_3B is dominated by D_3, condition **(R)** rules out A_3B. Player 2 will then know that, if her information set is reached, A_3T

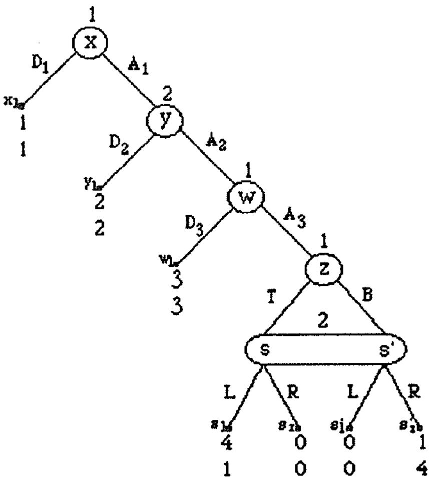

Figure 3: A more complex two-person game.

	D_2	A_2L	A_2R
D_1	1,1	1,1	1,1
A_1D_3	2,2	3,3	3,3
A_1A_3T	2,2	4,1	0,0
A_1A_3B	2,2	0,0	1,4

Figure 4: The strategic form of the game in Figure 3.

has been played, therefore she will respond with L. Foreseeing this reasoning of player 2, player 1 should play A_3T. In the subgame G' starting at node w, the equilibrium profile (D_3B, R), though subgame perfect, is ruled out by a forward induction argument. The equilibrium (A_1D_3, A_2R) thus violates all three conditions **(R)**, **(BI)** and **(ID)**.

Consider now equilibrium (A_1A_3T, D_2). What happens if a deviation occurs at node y? Condition **(R)** suggests that, by deviating, player 2 expects a higher payoff than what she gets by playing D_2. So it must be the case that, by deviating from the equilibrium path, player 2 is signaling that she will play R in the battle of the sexes. In which case it would be better for player 1 to play D_3. However, this reasoning is fallacious, since condition **(R)** must be applied iteratively. If rationality is common knowledge, it must also be common knowledge that, at node w, player 1 will not play D_3 but A_3T. Hence 2's best reply is L. Since it must be common knowledge that at node w player 1 will play A_3T, it follows that in the subgame G' starting at node y player 2 will choose D_2. The equilibrium (A_1A_3T, D_2) survives iterated application of condition **(R)**, and in addition it also satisfies **(BI)** and **(ID)**.

Given the above argument it follows that, if rationality is common knowledge, deviation A_2 should never be observed. Thus if the forward induction refinement has the advantage of making off-equilibrium beliefs consistent with common knowledge of rationality, its drawback is that it does not provide a complete theory of the game: Unexpected deviations cannot happen. One way to address this issue is to *complete* the theory with a model of belief revision, but in this paper we wish to take a different path.

Instead of computing the Nash equilibria for the game and then test them against deviations, we model how the players themselves reason to an equilibrium. The nature of such equilibria will obviously be a function of the way in which we characterize the players and their information. In the next section, we provide the players with a theory of the game T (a set of axioms) that embodies a simple rationality condition. We show that T leads to iterative elimination of dominated paths, that is, T generates an automatic decision procedure for the extensive form corresponding to iterated elimination of dominated strategies in the strategic form. Through iterated elimination of dominated paths we obtain a set (hopefully, a singleton) of undominated branches, and prove that each such branch corresponds to a Nash equilibrium for the game. Moreover, all equilibria thus obtained satisfy the criteria **(R)**, **(BI)** and **(ID)**. In the last section, we show that for T to be a complete theory it is sufficient to interpret some of its axioms as default rules.

3 The Theory

In general, a finite, extensive form game of imperfect information G is represented by a finite tree, having an arbitrary branching factor, equipped with two functions p and I. Function $p : G \rightarrow \{1, \ldots, k\}$ assigns a player i (for $0 < i \leq k$) to each node, while $I : G \rightarrow \mathcal{P}(G)$ assigns to each node the information set to which that node belongs. The branching factor of the tree is supposed to represent the number of choices available to each player at each node. In order to make things interesting, p is also assumed to be *non-injective*, thereby ensuring that at least one player gets to move more than once. We shall lay the following constraints on I, namely that (i) the sets it assigns are pairwise disjoint, (ii) that nodes belonging to the same information set are assigned to the same player. Payoffs at the terminal nodes (leaves) of the tree are represented by *real-valued vectors*, whose i-th projections (for $0 < i \leq k$) represent the payoff for player i at that node.

However, there is nothing conceptual to gain in representing such generality, while there is much to lose in notational perspicuity. All the points that we want to make can be made equally well for a restricted class of games. Consequently, we make the following simplifying assumptions. We will assume only two players that move in a pre-determined order (with a player possibly moving more than once in a row). Accordingly, payoffs at the leaves are represented by *pairs* of real values. It will be convenient to introduce a function $q : G - \{a\} \rightarrow \{1, 2\}$ (where a is the root of G), that for each node x other than the root gives the player that moves at the *previous* node.

We will also restrict ourselves to games represented by *balanced binary* trees, i.e., games in which each player has precisely two choices at each node and all branches have the same length. Conventionally, the two choices are referred to as "moving left" and "moving right." The trees are assumed to be "balanced," i.e., such that all branches have the same length: any unbalanced tree can be turned into a balanced one by adding nodes that are redundant from the point of view of the game (because they all lead to the same payoff vector). Similarly, we want information sets to be comprised of nodes that have the same distance from any leaf: in fact, information sets can be re-arranged in such a way that they contain only nodes of the same level in the tree. This can be accomplished as follows: if $x' \in I(x)$ is a node having the lowest level (i.e., the greatest distance from the root) among nodes in $I(x)$, when re-balancing the tree we put in the same information set all the nodes of the same level as x' that descend from nodes in $I(x)$ in the original tree.

CONVENTION Assume two players, 1 and 2, of whom player 1 is assumed to move first, so that the root of the tree represents a choice for 1. In what follows, a always denotes the *root* of G. Call a node *final* if it is non-terminal, but both its children are terminal. We write $x \sim y$ if x and y are "siblings," i.e., they are immediate successors of the same node. For any node x we denote by x_{r} and x_{l} its right- and left-hand successors, respectively. Moreover, by a *path* we mean a possibly empty sequence of nodes, each one of which is the successor of the previous one and the last of which is a leaf. A maximal path is called a *branch.* If x is a leaf, $\pi(x)$ represents the associated payoff vector, and if s is a path of length k, we write $\pi(s)$ instead of $\pi((s)_k)$. Also, we write, e.g., $\pi(x), \pi(y) > \pi(z)$ to mean $\pi(x) > \pi(z)$ and $\pi(y) > \pi(z)$. We use x, y, z as variables for nodes and s, t, u as variable for paths. If $b = \langle b_1, \ldots, b_k \rangle$ is any sequence and $i = 1, \ldots, k$, we set $(b)_i = b_i$. If s and t are sequences of nodes, their concatenetion is denoted by $s * t$.

Definition 3.1 *If P is a set of paths and x a node, define:*

$$\begin{aligned} \max_i(P) &= \{s \in P : \forall s' \in P\, (\pi(s))_i \geq (\pi(s'))_i\}; \\ \mathsf{Rest}(P, x) &= \{s \in P : (s)_0 \textit{ is a successor of } x\}. \end{aligned}$$

Intuitively, if P is a set of paths, $\max_i(P)$ returns the path(s) that maximize the utility for player i, i.e., those paths such that the payoff vectors associated with their leaves have a largest i-th component. Consider for instance the set of paths from Figure 1:

$$P = \{\langle y, y_{\mathrm{l}} \rangle, \langle y, y_{\mathrm{r}} \rangle, \langle y', y'_{\mathrm{l}} \rangle \langle y', y'_{\mathrm{r}} \rangle\};$$

Then $\max_2(P)$ returns $\langle y, y_{\mathrm{l}}$, since this is the one that gives highest payoff to player 2. On the other hand, $\mathsf{Rest}(P, x)$ returs those paths in P that begin with a node that descends from x. This is necessary, since there is no guarantee that if x and y are in the same information set then $x \sim y$.

We are ready to give the definition of the function π^* that associates with each information set the set of all paths undominated at that information set, from the point of view of the player whose turn it is to move at that set. The definition of $\pi^*(x)$ is by recursion on the level of x. This is sound, since all the nodes in $I(x)$ are assumed to be of the same level.

Suppose x to be a final node (i.e., a lowest non-terminal node). Suppose x belongs to an information set at which player i is called upon to move, and for notational convenience, let

$$\begin{aligned} S^* \quad = \quad & \{\langle x', y\rangle : x' \in I(x) \wedge y \text{ is a child of } x' \wedge \\ & \neg\exists z(x' \sim z \wedge (\pi(z_{\mathrm{l}}))_{q(x)} > (\pi(x'_{\mathrm{l}}))_{q(x)} \wedge (\pi(z_{\mathrm{r}}))_{q(x)} > (\pi(x'_{\mathrm{r}}))_{q(x)}\}. \end{aligned}$$

Intuitively, S^* is the set of all length-two paths starting from a node x' in $I(x)$ that are *not locally dominated* (from the point of view of the *previous* player) by a sibling: node x' is locally dominated by z if from the point of view of the previous player moving left at z is strictly better than moving left at x', and moving right at z is strictly better than moving right at x'. Nodes that are dominated in the above sense are better left out of consideration, since the player whose turn it is to move at the common parent of x and x' can be assumed never to choose a dominated path. (This is where assumption (**R**) comes into play: eliminating these paths embodies the essence of forward induction.) We can then set :

$$\pi^*(I(x)) = \bigcup\{\max_i(\mathsf{Rest}(S^*, x') : x' \in I(x)\}.$$

That is, for each node x' in $I(x)$, π^* returns the best paths that are not dominated from the other player's point view.

If, on the other hand, x is not final, let x_{r} and let x_{l} be its children, and suppose as before that player i is called upon to play at x_{r} and x_{l} (so that player $q(x_{\mathrm{r}}) = q(x_{\mathrm{l}})$ is called upon to play at x). By inductive hypothesis both $\pi^*(I(x_{\mathrm{r}}))$ and $\pi^*(I(x_{\mathrm{l}}))$ are defined. Again, given a node y, we want to say that y is locally dominated for i if there is a node y' such that $y \sim y'$, and all paths beginning at y'_{r} are strictly better (in the sense of leading to better payoffs for i) than all paths beginning at y_{r}, and all paths beginning at y'_{l} are strictly better than all paths beginning at y_{l}. So let $\mathsf{Dom}(x, i)$ be understood as follows:

$$\begin{aligned} & \mathsf{Dom}(x, i) \iff \\ & \quad \exists y[x \sim y \wedge (\forall s \in \pi^*(I(y)))(\forall s' \in \pi^*(I(x))) \\ & \qquad [(s)_1 = y \wedge (s')_1 = x \wedge ((s)_2 = y_{\mathrm{r}} \leftrightarrow (s')_2 = x_{\mathrm{r}}) \\ & \qquad\quad \rightarrow (\pi(s))_i > (\pi(s'))_i]]. \end{aligned}$$

(The condition $((s)_2 = y_{\mathrm{r}} \leftrightarrow (s')_2 = x_{\mathrm{r}})$ ensures that we only compare "left moves" with "left moves," and "right moves" with "right moves.") Now define, for each $x' \in I(x)$, a set of paths $S^*_{x'}$, as follows:

$$\begin{aligned} S^*_{x'} \quad = \quad & \{\langle x'\rangle * s : s \in \mathsf{Rest}(\pi^*(I(x'_{\mathrm{r}})), x) \wedge \neg\mathsf{Dom}(x', q(x))\} \cup \\ & \{\langle x'\rangle * s : s \in \mathsf{Rest}(\pi^*(I(x'_{\mathrm{l}})), x) \wedge \neg\mathsf{Dom}(x', q(x))\} \end{aligned}$$

This is the set of all paths descending from x' that are not locally dominated from the point of view of the other player: the player whose turn it is to move at x will then choose, *among these*, the ones that are best from her own point of view. This is accomplished by setting:

$$\pi^*(I(x)) = \bigcup\{\max_i(\mathsf{Rest}(S^*_{x'}, x') : x' \in I(x)\}.$$

Let us see how this definition works by taking up the game of Figure 3. First we consider the nodes in $I(s)$. This is an information set for player 2, with 1 moving at z. Since none is locally dominated, π^* returns the best path starting at s and the best path starting at s': these are $\langle s, s_{\mathrm{l}}\rangle$ and $\langle s', s'_{\mathrm{r}}\rangle$, which give payoffs of 1 and 4, respectively. Now we compute $\pi^*(\{z\})$: first we check if any path in $\pi^*(I(s)) = \pi^*(I(s'))$ is locally domainated; indeed, s' is locally dominated by w_{l}, so path $\langle z, s, s_{\mathrm{l}}\rangle$ survises, since it is the one affording the best payoff for 1; the operation of doing $\max_1$ is now idle, since it is apllied to only one path, which is returned as output. Similarly, player 1 chooses again at node w: he computes $maxi_1(P)$, where P contains the paths $\langle w, s, s_{\mathrm{l}}, z\rangle$ and $\langle w, w_{\mathrm{l}}\rangle$: obviously, this gives $\langle w, z, s, s_{\mathrm{l}}\rangle$. it is then 2's turn to move: he has to choose between $\langle y, w, z, s, s_{\mathrm{l}}\rangle$ and $\langle y, y_{\mathrm{l}}\rangle$ and he clearly chooses the latter. Since this is also a best choice for 1's point of view at node x, the path $\langle x, y, y_{\mathrm{l}}\rangle$ is the only path returned by $\pi^*(I(x))$.

It is important here two notice to things. First, that the procedure would have given precisely the same outcome if the order in which the players move in the final "battle of the sexes" had been reversed. Moreover, in case all information sets are singletons, the outcome of the procedure coincides with backwards induction (this can be easily checked for instance in the case of the game of Figure 3).

Our next step is to prove that the paths in $\pi^*(I(x))$ are all Nash equilibria. In order to do so, we need to define the analog of the idea of Nash equilibrium for extensive form games. A branch through the tree corresponds not to one but to *two* strategies, one for each player. We thus have to compare the payoff of a given path (for a given player) with the payoffs of all other paths that embody alternative strategies, keeping fixed, however, the elements of the original path that correspond to a strategy for the other player.

We now proceed to capture this formally as follows. By a *move* we understand a length-two sequence of nodes such that the second is a child of the first. A *strategy*, in turn, is a set of moves, some of which might never be played if the strategy is chosen.

We start from a path s and partition it into two sets $M_1(s)$ and $M_2(s)$ corresponding to the moves of the two players: M_1 is the set of all length-two paths $\langle x, y\rangle$ such that $\langle x, y\rangle$ is a subpath of s, and player 1 moves at x. Similarly, M_2 is the set of all length-two paths $\langle x, y\rangle$ such that $\langle x, y\rangle$ is a subpath of s, and player 2 moves at x. We then expand $M_i(s)$ to a set S that is (1) "move-uniform," and moreover (2) contains a response to each possible move of $3 - i$. To say that S is move-uniform means that it satisfies the following condition: if $\langle x, y\rangle \in S$, $x' \in I(x)$ and y' is the left- or right-hand child of x' according as y is the left- or right-hand child of x, then $\langle x', y'\rangle \in S$, too. To say that it contains a response to each possible move of $3 - i$ means that it satisfies the following condition: if $\langle x, y\rangle \in S$ and $3 - i$ moves at y and z is a child of y, then there is exactly one move $\langle z, u\rangle$ that belongs to S. It is clear that $M_i(s)$ can always be extended (non-deterministically) to a set S satisfying the above two conditions: S corresponds to a strategy in normal form games. So define a set of moves S to be a *strategy* for i if it is a minimal set of moves containg $M_i(s)$ for some branch s and satisfying (1) and (2).

Given a strategy S_1 for player 1 and a strategy S_2 for player 2, there is at most one branch s such that all and only its length-two segments are contained in $S_1 \cap S_2$, in which case (by slightly abusing the language) we will write $s = S_1 \cap S_2$. Define $\mathbf{u}_i(S_1, S_2)$, the payoff for player i of playing strategy S_1 against strategy S_2 as $(\pi(s))_i$, if there is a unique branch s contained in $S_1 \cap S_2$, and set $\mathbf{u}_i(S_1, S_2) = -\infty$ otherwise.

A pair of strategies (S_1, S_2) (for players 1 and 2, respectively), is a Nash equilibrium if and only if:

- for any strategy S' for 1, $\mathbf{u}_1(S', S_2) \leq \mathbf{u}_1(S_1, S_2)$; and
- for any strategy S' for 2, $\mathbf{u}_2(S_1, S') \leq \mathbf{u}_2(S_1, S_2)$.

THEOREM 3.1 *Let $s \in \pi^*(I(a))$, where a is the root of the tree, and let S_1, S_2 be any two strategies such that $s = S_1 \cap S_2$. Then (S_1, S_2) is a Nash equlibrium.*

Proof. Suppose for contradiction that (S_1, S_2) were not a Nash equilibrium. Then one of S_1, S_2 is dominated by some strategy S' (from the point of view if player i). Suppose S_1 is dominated by S' for player 1. Of course, this can only happen if the two strategies intersect to give a branch, otherwise $u_1(S', S_2) = -\infty$. So there must be a terminal node x such that S_1 contains $\langle x, x_r \rangle$, S' contains $\langle x, x_l \rangle$, and $(\pi(x_r))_1 < (\pi(x_l))_1$. But this means that s is a dominated branch, and hence $s \notin \pi^*(I(a))$, against hypothesis. ∎

We now proceed to give a first-order theory T that, employing function π^*, allows us to predict the players' behavior. Since the definition of π^* is formalizable in Primitive Recursive Arithmetic, we shall assume that T contains enough arithmetic to carry out that definition. Moreover, T will have to contain axioms describing the tree representing the game and the structure of the payoffs at the leaves. We shall assume that all these "structural" axioms have been specified, and proceed to give our "behavioral" axioms.

First of all, we want to say that if a certain non terminal node is reached, then the player whose turn it is to move will choose exactly one of the possible moves. We introduce predicates $L(x)$ and $R(x)$ with the intended meaning that the player (whose turn it is to move) moves left and, respectively, right at node x. If a is the root of the tree, we introduce an axiom

$$R(a) \leftrightarrow \neg L(a). \tag{1}$$

Moreover for each non-terminal node x other than the root, we proceed as follows: let $x_1, \ldots, x_{n+1} = x$ be the nodes on the path from the root to x, and let Q_i be R or L according as x_{i+1} is reached by moving left or right at x_i. Then we introduce an axiom saying that

$$Q_1(x_1) \wedge \ldots Q_n(x_n) \rightarrow (R(x) \leftrightarrow \neg L(x)). \tag{2}$$

Next, we introduce an individual constant s representing (a suitable coding) of the set of undominated paths. We introduce an axiom to the effect that such a set is obtained by our procedure:

$$\mathsf{s} = \pi^*(I(a)), \tag{3}$$

where again a is the root of the game. Finally, we introduce axioms saying that players only move along the undominated paths: let y be any non terminal node and suppose it has height (= number of predecessors) k. Then we have the axioms:

$$L(y) \rightarrow \exists s \in \mathsf{s}\,((s)_k = y_l); \tag{4}$$

$$R(y) \rightarrow \exists s \in \mathsf{s}\,((s)_k = y_r). \tag{5}$$

This completes our specification of T, which will then comprise (1)–(5) as behavioral axioms.

4 Towards a Complete Theory

We now indicate how the previous theory T can be modified to handle real deviations from equilibrium (i.e., those deviations that cannot be consistently be interpreted as "signals"). We are going to interpret T as a *default theory* $T' = (W, D)$, where W is a set of first-order axioms and D a set of normal defaults. A normal default is a weak inference rule of the form $\varphi \rightsquigarrow \psi$, interpreted as saying "if φ is known, and ψ is consistent, infer ψ" The sense in which ψ has to be consistent in order for it to be inferable is made precise in Default Logic; see Reiter [9] for details.

We have to specify W and D. In our theory, W comprises all the "structural axiom," i.e., whatever arithmetic is necessary to describe the game and compute π^*, along with a suitable coding of the game and associated payoffs. As before, we will leave this unspecified. Moreover, W will contain all formulas of the form (1) and (2).

On the other hand, D will specify the set of paths to be used in inferring the players' behavior. This set of paths will vary according as the node that has been reached in the game lies on or off the equilibrium paths.

Let a be the root, and $\top$ a propositional constant representing truth. Then we have a default to the effect that at the beginning of the game we use the equilibrium paths provided by $\pi^*(I(a))$:

$$\top \rightsquigarrow \mathbf{s} = \pi^*(I(a)); \tag{6}$$

for any other node x, let $x_1, \ldots, x_{n+1} = x$ be the nodes on the path from the root to x, and let Q_i be R or L according as x_{i+1} is reached by moving left or right at x_i. Then we introduce a default of the form:

$$Q_1(x_1) \wedge \ldots Q_n(x_n) \rightsquigarrow \mathbf{s} = \pi^*(I(x)). \tag{7}$$

This completes our specification of T'. This theory is *complete* in the sense that it can be augmented with information to the effect that a real deviation has taken place without becoming inconsistent. Moreover, it still allows us to say something about the game *after* a deviation, in the sense that it will still have an extension (in the sense of Reiter [9]) according to which all moves following a deviation still take place along of the paths that are undominated in the subgame whose root is represented by the node at which the deviation has taken place. In this sense, T' embodies a principle of *local rationality*, non dissimilar to the one in Bicchieri & Antonelli [5].

References

[1] R. Aumann, *Backwards Induction and Common Knowledge of Rationality*, mimeo, University of Jerusalem, 1993.

[2] C. Bicchieri, *Self Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge*, **Erkenntnis** 30 (1989), pp. 69–85.

[3] C. Bicchieri, *Knowledge-dependent Games: Backward Induction*, in Bicchieri & Dalla Chiara, **Knowledge, Belief, and Strategic Interaction**, Cambridge University Press, Cambridge 1992.

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

[5] C. Bicchieri & G.A. Antonelli, *Game-Theoretic Axioms for Bounded rationality and Local Knowledge*, paper given at the Nobel Symposium on Game Theory, Björkborn, Sweden, June 1993.

[6] E. Kohlberg & J.F. Martens, *On the Strategic Stability of Equilibria*, **Econometrica** 54 (1986), pp. 1003–37.

[7] D. Kreps and R. Wilson, *Sequential Equilibria*, **Econometrica** 50 (1982), pp. 863–94.

[8] R.B. Myerson, *Credible Negotiation Statements and Coherent Plans*, **Journal of Economic Theory**, v. 48 (1989), pp. 264–303.

[9] R. Reiter, *A Logic for Default Reasoning*, **Artificial Intelligence**, v. 13 (1980), pp. 81–132.

[10] R. Selten, *Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit*, **Zeitschrift für die gesampte Staatswissenschaft** 121 (1965), 667–89.

[11] R. Selten, *Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games*, **International Journal of Game Theory**, v. 4 (1975), pp. 25–55.

[12] R. Stalnaker, *Knowledge, Belief, and Counterfactual Reasoning in Games*, forthcoming in the proceedings of the second Castiglioncello conference 1992, edited by C. Bicchieri and B. Skyrms.