PHILOSOPHICAL TOPICS
VOL. 21 NO. 1, SPRING 1993

Counterfactuals, Belief Changes, and Equilibrium Refinements[1]

Cristina Bicchieri
Carnegie Mellon University

INTRODUCTION

It is usually assumed in game theory that agents who interact strategically with one another are rational, know the strategies open to other agents as well as their payoffs, and, moreover, have common knowledge of all the above. In some games, that much information is sufficient for the players to identify a "solution" and play it. The most commonly adopted solution concept is that of Nash equilibrium. A Nash equilibrium is defined as a combination of strategies, one for each player, such that no player can profit from a deviation from his strategy if the opponents stick to their strategies. Nash equilibrium is taken to have predictive power, in the sense that in order to predict how rational agents will in fact behave, it is enough to identify the equilibrium patterns of actions. Barring the case in which players have dominant strategies, to play her part in a Nash equilibrium a player must believe that the other players play their parts, too. But an intelligent player must immediately realize that she has no ground for this belief. Take the case of a one-shot, simultaneous game. Here all undominated strategies are possible choices, and the beliefs supporting them are possible beliefs, even if this game has a unique Nash equilibrium. The beliefs that support a Nash equilibrium are a subset of the beliefs that players may plausibly have, but

nothing in the description of the game suggests that players *will* in fact restrict their beliefs to such a subset.

We only know that if each player plays her part in the equilibrium, and each expects the others to play their parts, each player behaves correctly in accordance with her expectations and each has confirmed the others' expectations. In other words, the beliefs that support a Nash equilibrium are always correct beliefs, that is, they are mutually consistent. It may seem that equilibrium play would be guaranteed were the players to have common knowledge of their beliefs.[2] But it is easy to think of examples in which the players start out with mutually inconsistent beliefs, and common knowledge of their respective beliefs will only generate deliberational cycles.[3] To attain predictability, we thus need to specify some mechanism through which beliefs become correct (i.e., mutually consistent). For example, in the absence of direct knowledge of one another's beliefs, the players will have to infer them from observed actions, or at least they must be able to restrict the class of beliefs an opponent may plausibly entertain. One way to restrict the class of possible beliefs is to consider only beliefs that are plausible or rational in a substantive sense. In the normal form representation of a game, however, belief rationality is just a matter of internal consistency. The focus is on how a belief coheres with other beliefs that one holds, while it is irrelevant how well founded it is.

A substantive interpretation of belief rationality involves assessing whether a belief is justified, and one way to do it is by identifying those beliefs that are the outcome of a rational process of belief formation. The extensive form representation of the game, by specifying the causal structure of the sequence of decisions and the information available at each decision point, is the proper setting for modeling the process of belief formation. I shall also argue that a satisfactory theory of belief formation must tell how players would change their beliefs in various hypothetical situations, as when confronted with evidence inconsistent with formerly accepted beliefs. The theory of belief revision I propose is based on a principle of minimum loss of informational value. The informational value of a proposition reflects its predictive and explanatory potential, and this is a function of what players want to explain and predict. The criterion of informational value presented here induces a complete and transitive ordering of the sentences contained in a belief set. So a player who revises her beliefs rationally, in accordance with that theory, will eliminate first those beliefs that have low informational value.

To predict the outcome of someone's belief-revision process one has to know, among other things, the revisor's rules for belief revision, as well as her explanatory and predictive interests. I shall assume that the rules for belief revision, as well as the criterion of informational value, are shared by the players. The rules for belief revision specify a criterion of equilibrium

selection; if this criterion identifies a unique equilibrium as the solution for the game, then players who have common knowledge of rationality, of the shared rules for belief revision, and of the shared criterion of informational value can identify their equilibrium strategies. In this case, players' beliefs will be both correct and common knowledge.

The case of a unique Nash equilibrium is straightforward, since whenever there is a unique solution for the game this solution is a fortiori the one selected by belief revision. The interesting case is that in which there are multiple Nash equilibria, some of which might be implausible in that they involve "risky" 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 multiple equilibria. These refinements correspond to different ways to check the stability of a Nash equilibrium against deviations from equilibrium play. The players are supposed to agree to play a given equilibrium and then ask what would happen were they (or their opponents) to play an off-equilibrium strategy. If the players decide that they would play their part in the equilibrium even in the face of deviations, then that equilibrium is stable (or plausible). Stability, however, is a function of how a deviation is being interpreted. A player may deviate from an expected equilibrium because she is irrational, because she made a mistake, or perhaps because she wants to communicate something to her opponents. An equilibrium that is stable under one criterion may cease to be stable under another. Different refinements propose different interpretations for deviations, and there is no clear sense of how to judge their plausibility and, when two or more interpretations are possible, how to rank them.

When facing another player's deviation, a player has to modify her beliefs, but the current refinements of Nash equilibrium fail to specify criteria of belief revision that would restrict players' explanations of deviations (off-equilibrium beliefs) to a "most plausible" subset. In this paper, I first introduce a set of simple and plausible restrictions that any off-equilibrium belief should satisfy. I then show that such plausible explanations of deviations are the result of a rational process of belief revision; that is, a process of belief revision that minimizes the loss of useful information. The theory of belief revision I propose succeeds in generating a ranking of interpretations of deviations, hence it also generates a ranking of the most common refinements. When several interpretations are compatible with a deviation, the one that requires the least costly belief revision (in terms of informational value) will be preferred. A consequence of the theory of belief revision presented here is that it leads players to interpret deviations, whenever possible, as intentional moves of rational players, thus providing a strong theoretical justification for forward induction arguments.

To model belief formation, it is useful to consider the dynamic structure of games, the order in which players move and the kind of information they have when they have to make a choice. Briefly, the extensive form of a game specifies the following information: a finite set of players i = 1, . . . n, one of which might be nature (N); the order of moves; the players' choices at each move and what each player knows when she has to choose; the players' payoffs as a function of their moves; finally, moves by nature correspond to probability distributions over exogenous events. The order of play is represented by a game tree T, which is a finite set of partially ordered nodes $t \in T$ that satisfy a precedence relation denoted by "<".[4] The information a player has when he is choosing an action is represented using information sets, which partition the nodes of the tree. Since an information set can contain more than one node, the player who has to make a choice at an information set that contains, say, nodes t and t' will be uncertain as to which node he is at.[5] If a game contains information sets that are not singletons, the game is one of *imperfect information,* in that one or more players will not know, at the moment of making a choice, what the preceding player did. Finally, the games I shall consider are all games of *perfect recall,* in that a player always remembers what he did and knew previously.

Figure 1 is a simple example of a two-players' extensive form game. In this particular game there is an initial starting point, or initial node, at which player 1 has to move. If he chooses L, the game ends with both players getting a payoff of 1. If 1 chooses R instead, it is player 2's turn to move, and she can choose between actions l and r. If the choices of R and l are taken, then both players net -1. If instead R and r are chosen, player 1 gets 2 and player 2 gets 0.

The game has two Nash equilibria in pure strategies, (L,l) and (R,r). In the normal form representation of the game (figure 2), there is no way to predict with confidence which pair of actions will be chosen by the players, at least if one remains agnostic about the beliefs of both players.

The equilibrium (L,l) is not implausible if player 2 believes that L is played and, in turn, player 1 believes that 2 selects l, even if strategy l is weakly dominated by r.[6] Nash equilibria are often equated with self-enforcing agreements. That is, if the players agree to play a given pair of strategies and no one has an incentive to deviate from his agreed-upon strategy (provided he

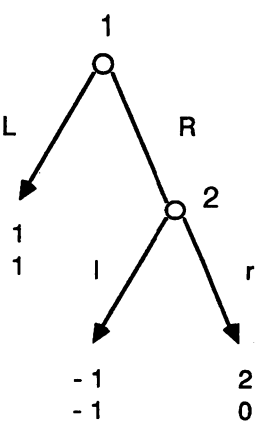

FIGURE 1

believes the opponent is sticking to the agreement), then that pair of strategies is a Nash equilibrium. Suppose the two players of the game in figure 2 can meet before playing and agree to play the strategy profile (L,l). Is (L,l) a self-enforcing agreement? Yes, if each player believes that the other will stick to his part of the agreement. But what should be also asked is whether the agreement is reasonable. An agreement is unreasonable if a player cannot justify the claim that it will be honored except by adopting unreasonable expectations about what his opponent is likely to do or her reasons for doing it.

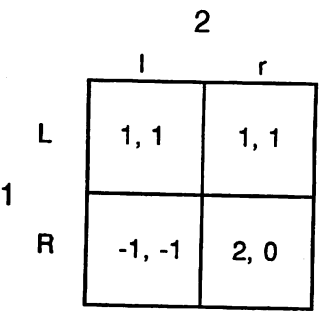

FIGURE 2

We are supposing that player 2 tells player 1 that, come what may, she will play l. So it is understood that, were 1 to play R, both would net -1. Now consider the extensive form game of figure 1. If player 1 were to play R instead of L, would 2 stick to the original agreement and respond with l? Clearly if 2 were to reach her decision node, she would choose the payoff maximizing action r. Since player 1 knows that 2 is rational, he should never play L, for he will always get a higher payoff by playing R instead. It follows that even if in the normal form the equilibrium (L,l) is supported by a set of consistent beliefs, it is clearly unreasonable in the extensive form representation of figure 1. There, it involves the irrational expectation that player 2, once she is called upon to play, will still choose to play l.

The example is meant to show that what constitutes a reasonable commitment to play a Nash equilibrium is affected by what one supposes will be another's action out of equilibrium, i.e., what reaction one expects if one deviates from the equilibrium path. Note that a Nash equilibrium does not involve any prescription (or restriction) on out-of-equilibrium behavior. The only restrictions imposed are those on equilibrium actions (i.e., that they are best replies). At first sight, the concern with out-of-equilibrium behavior seems paradoxical: if both players play a Nash equilibrium, actions which lie out of the equilibrium path are never performed, since by definition the information sets at which they would be chosen are never reached. Then of course *any* action out of the considered equilibrium path is admissible, since it remains an intention that will never be carried out. Indeed, if the equilibrium (L,l) is played, it does not really matter what player 2 does, since she will never have to choose.

One traditional justification for Nash equilibrium is that it is an agreement that holds up despite the absence of an enforcement mechanism. When multiple equilibria are present, an important step toward predictability is to rule out those equilibria that are not robust to potential deviations, since they constitute agreements that we would not expect rational players to hold. To

illustrate the point, suppose you and I agree to meet in an hour at the campus cafeteria. Since you have every reason to expect me to be there, any question about what you would do if I were not to be there at the appointed time seems futile. But assume now that you threaten me by saying that if I am even five minutes late, you will immediately leave the cafeteria without eating. From the viewpoint of our being there on time, what you would do under different circumstances is irrelevant and, more to the point, all sorts of behaviors are admissible. On closer scrutiny, however, what you would do if I were not there on time *does* matter to my decision of whether to hurry or not. We know each other very well, so I know that in an hour you will be hungry, and since there is only that cafeteria around, your threat is hardly believable. Therefore I can take my time. These considerations are obviously relevant to our original agreement. Since we both know that your threat is not credible, we may still agree to meet at the cafeteria, but be flexible as to the amount of time either of us might spend waiting for the other.

Note that the original agreement-plus-threat is an equilibrium, since if we both believe the other will fulfill the terms of the agreement, neither of us has an incentive to deviate. Our beliefs are both internally consistent and correct (indeed, each of us does what she is expected to do), but are they plausible? If we do not find them plausible, neither is the agreement that they support. To establish whether the original agreement is sensible, we have to ask what would happen out of equilibrium (that is, in case one of us "deviates" by breaking the agreement). In the present case, considering the hypothetical situation (from the viewpoint of the original agreement) in which I am late leads us to conclude that you will still go into the cafeteria and eat, and therefore it rules out an agreement in which the latecomer is penalized. In other words, we check the reasonableness of an agreement by considering what would happen if one or more of the parties were to deviate from it. Hence, one should ask not only whether it is sensible to honor an agreement were the other party to honor it, but also whether the other party would find it in her interest to honor the agreement were one to break it. This reasoning highlights the importance of the credibility of the threats supporting an equilibrium; if our agreement is based upon my threat to retaliate if you do not perform a given action, I'd better make sure that you believe my threat. That is, it must be evident to both of us that I will honor my end of the agreement (and thus punish you) in case you defect.

BACKWARD INDUCTION

The methodology employed here is more complex than that used to verify that an agreement is a Nash equilibrium. In the latter case, one asks whether it would be in one's interest to deviate from the prescribed course of action

in case everybody else honors the agreement. In our example, a player asks whether the other player would honor the agreement were he to break it. In figure 1, for example, player 1 may wonder what would happen if, after agreeing to play the equilibrium (L,l), he were to deviate and play R instead. Player 1 wants to know whether it is sensible to deviate from the intended course of action, given the foreseen reaction of the opponent. In the simple game of figure 1 it is easy to predict that player 2, being rational, will respond to R with strategy r. The problem is that there are many games in which it is not so obvious what the opponent's reaction to a deviation would be. It all depends on how one's deviation is explained. A first step in deciding whether a Nash equilibrium is a sensible agreement thus consists in placing restrictions on out-of-equilibrium actions, a step which corresponds to restricting the set of possible explanations for those actions. Such explanations constitute what I call "out-of-equilibrium beliefs." Out-of-equilibrium beliefs are the beliefs (attributed to herself and to other players) that a player *now* thinks would explain a given off-equilibrium choice.

If restricting out-of-equilibrium beliefs is a necessary step in deciding whether an equilibrium is sensible, and thus in predicting behavior as precisely as possible, one may wonder whether the same goal would be accomplished by considering only those equilibria that do not involve irrational (i.e., dominated) actions. The rationale for this proviso is as follows: since off-equilibrium choices are relevant only when they affect the choices along the equilibrium path, it seems 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 figure 1, player 1 knows that player 2 is rational and since rational choice is undominated, he knows that 2 will never play a dominated strategy if she were to reach her decision node; this consideration rules out the equilibrium (L,l) as a plausible self-enforcing agreement. Considering only undominated actions means that out-of-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 dominated strategy. In other words, it must be common knowledge that weakly dominated strategies will not be used. In many games, common knowledge of rationality is not even needed to rule out dominated strategies. In figure 1, for example, player 1 has to know that 2 is rational in order to predict her choice, but no further knowledge is needed on his part. And player 2, being the last one to choose, need not know anything about 1's rationality, since what happens before her

decision node is irrelevant to her choice, given that she has one. To decide whether a strategy is dominated is not always such a simple matter. In those games in which iterated elimination of dominated strategies applies, whether or not a strategy gets to be dominated may depend on one's beliefs about the opponent's choices and beliefs. That is, if we eliminate a number of (dominated) options for the opponent, this affects what is dominated for us. But in order to eliminate an opponent's dominated strategy, a player must know (or at least be reasonably certain) that the opponent is rational and, depending on the round of elimination, that several iterations of "He knows that I know that . . . he is rational" and "I know that he knows that . . . I am rational" obtain. This is why we say that successive elimination of dominated strategies involves more information than one round of elimination.

In this paper I am only considering extensive form games. How does successive elimination of dominated strategies work in such games? Or, to put it differently, how much information does a player need in order to decide that a given strategy is dominated? Consider the two-players' game form in figure 3.

Suppose it is optimal for each player to play "d" at every decision node. Then an optimal strategy for player 1 is to play "d at node x and d at node j," even if "play d at node j" is a recommendation he will never have to follow, given that he plays "d" at his first node and thus ends the game.[7] How does player 1 decide that playing "d" at node j is optimal? The decision is straightforward if the outcome of "d" is better than any possible outcome 1 might obtain by playing "a" (and leaving the choice to player 2 at node z). However, if one of player 2's successive choices (at node z) might get 1 a better payoff than playing "d" at j does, it would matter to player 1 what he expects that 2 would do at node z, were 1 to play "a" instead of "d" at node j. In conjecturing 2's future intentions, player 1 must consider that, if he has reached node j, this means that 2 did not choose "d" at node y; hence, what 1 believes 2 will do at node z depends on how he explains 2's choice of "a" at node y. Unless the outcome of playing"d" at node y is inferior to any

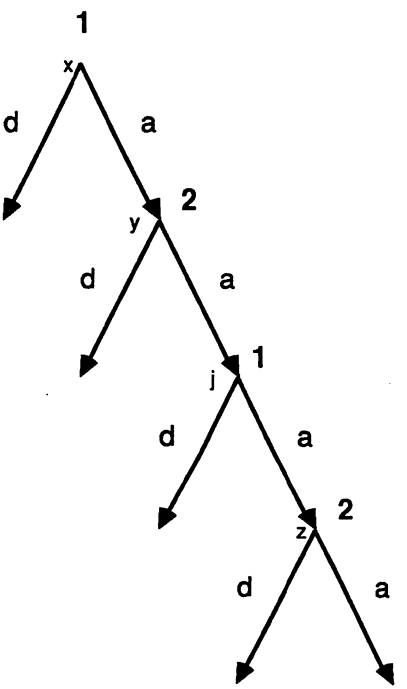

FIGURE 3

other outcome 2 might obtain by playing "a", her choice will depend on what 2 herself believes 1 would choose at node j, given that he played "a" at node x. So player 1's strategy at node j may have to include an assessment of the beliefs 2 has at node y regarding 1's future play. In this light, it becomes apparent that what constitutes an optimal choice for a player might depend upon his beliefs about the opponent's play (and beliefs). As an example, think of 2's choice at node z. Suppose that the payoff to 2 that comes from playing "d" is greater than what 2 gets by playing "a". If rational, 2 will certainly choose "d". Suppose further that were 2 to choose "a" at z, it would yield an outcome that 1 prefers to the outcome of choosing "d" at node j. Then 1's best reply to player 2's choice of "a" at z would be to play "a" himself at node j. Whereas "d" would be 1's best choice at node j if 2 is expected to play "d" at z. At node j, "d" dominates "a" for player 1 if he expects 2 to play "d" at node z, otherwise "a" dominates "d". It clearly matters to 1's decision that "d" dominates "a" at node j whether or not he knows that 2 is rational.

As I mentioned at the outset, condition (R) might even be stronger than necessary for most games. In fact, in finite, extensive form games of perfect information the number of levels of mutual knowledge of rationality that is sufficient for the players to infer a solution is finite, and it depends on the length of the game.[8] In such games, (R) guarantees that each player will play her part in the backward induction equilibrium. Backward induction in fact excludes implausible Nash equilibria, since it requires rational behavior at all nodes.

FORWARD INDUCTION

Up to now I have considered extensive form games of *perfect information.* In such games there are no simultaneous moves, and at each decision point it is known which choices have been previously made. In these games backward induction does two quite different things: a) it involves a computational method that, in the absence of ties, determines a single outcome, and b) it excludes all implausible Nash equilibria, since it requires rational behavior even in those parts of the tree that are not reached if the equilibrium is played. Using backward induction thus allows us to winnow out all but the equilibrium points that are in equilibrium in each of the subgames and in the game considered as a whole.[9] More generally, we may state the following backward 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 initial node of the game tree.

Coupling the conditions (R) and (BI) guarantees that unreasonable equilibria are ruled out, thereby leading to greater predictive power. In the game of figure 1, for example, (BI) rules out strategy l for player 2. Strategy l is a best reply to L, but it is not a best reply to R. Condition (R) requires beliefs to be consistent with common knowledge of rationality, where a definition of rationality includes admissibility (i.e., a player will not choose a dominated action).[10] Together with (BI), (R) implies that a self-enforcing Nash equilibrium must be consistent with deductions based on the opponent's rational behavior in the future. Future behavior, however, may involve out-of-equilibrium behavior, for when the equilibrium is played no further choices may take place. As I mentioned at the outset, out-of-equilibrium actions and beliefs need to be restricted to ensure predictability. Condition (R) provides such a restriction since it implies that out-of-equilibrium actions must be restricted to the set of undominated actions, so the only deviations that matter are those that can be interpreted as intentional choices of rational players.

Note that in the games considered thus far the same epistemic conditions ensure that deductions based on the opponent's behavior in the future (backward induction) agree with deductions based on the opponent's rational behavior in the past (forward induction). With backward induction, the fact that a node is reached does not affect what happens there. That is, we can ignore the earlier part of the tree in analyzing behavior at that node. With forward induction, on the other hand, deviations from an equilibrium are taken to be 'signals', intentional choices of rational players.[11] So if a node is reached, one asks why a deviation occurred, and one tries to give an explanation that is consistent with maintaining that the deviating player is rational. This is not the unique interpretation of deviations that makes them compatible with rational behavior, though. A deviation might be due to a mistake, or it might be possible that one of the players has an incorrect model of the game. These alternative explanations and their shortcomings will be discussed later. My concern in what follows is with the general applicability of criteria such as (R) and (BI) to different classes of extensive form games. Consider the game in figure 4.

This game is one of *imperfect information,* in that player 1, when it is his second turn to move, is unable to discriminate between z and z', i.e., he does not know what player 2 did before. The set {z, z'} is called the *information set* of player 1, and is denoted by a dotted line. The backward induction approach fails here, since at 1's information set there is no unique rational action; in z, player 1

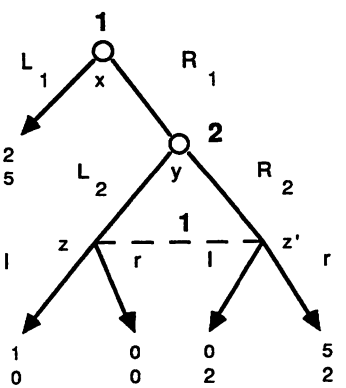

FIGURE 4

should play l and in z', he should play r. There is no way to define an optimal choice for player 1 at his information set without first specifying his beliefs about 2's previous choice. The backward induction algorithm fails because it presumes that such an optimal choice exists at every information set, given a specification of play at the successors of that information set. Even if backward induction is not defined in a game like the one in figure 4, the idea of working from the end of the game upwards can still be exploited. If there exist subgames, one can ask whether an equilibrium for the whole game induces an equilibrium in every subgame. This suggests that condition (BI) can still apply even when the backward induction procedure is not defined.

A refinement of Nash equilibrium that applies condition (BI) to games of imperfect information is the *subgame perfect equilibrium.*[12] A subgame perfect equilibrium is a Nash equilibrium such that the strategies when restricted to any subgame form a Nash equilibrium of the subgame. In figure 4, the subtree starting at y constitutes a game of its own. Since the game is non-cooperative, there are no binding commitments, hence behavior at node y is only determined by what comes next. At node y player 2 will choose R_2, which leads to a better payoff whatever 1 does. Knowing that 2 is rational, 1 will assign probability 1 to z', and thus play r; (rR_2) is the only equilibrium for the subgame starting at node y, hence (R_1rR_2) is the only sensible (i.e., subgame perfect) equilibrium, whereas ($L_1l\,L_2$), though a Nash equilibrium, cannot induce an equilibrium in the subgame starting at y.

Subgame perfection succeeds in excluding certain types of equilibria by defining a subclass of equilibria that all satisfy the (BI) requirement, but it may fail to rule out unreasonable equilibria when there are no subgames. Moreover, even when there are subgames, subgame perfection may be too weak a criterion, in that (BI) may not lead to a definite prescription of play. Consider the game in figure 5.

In the subgame starting at y, player 2 has no dominant strategy, so player 1 can assign any probability to z and z'. Both (L_2l) and (R_2r) are Nash equilibria of the subgame. Hence ($R_1l\,L_2$) and ($L_1r\,R_2$) are both subgame perfect even if, as I argue below, one would think that ($R_1l\,L_2$) is more plausible. Here the (BI) condition does not help in deciding what to do, but condition (R) does. Since by assumption rationality is common knowledge and R_1r is dominated by L_1 (that is, R_1r yields at best a payoff of 1, while L_1 yields 2), it is common knowledge that 2 does not expect 1 to

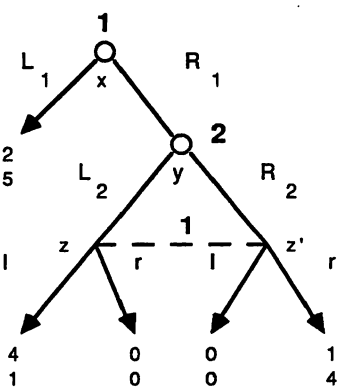

FIGURE 5

play R_1r. Therefore, it is common knowledge that, since 1 would never choose R_1r, if 1 picks R_1, he must be planning to follow that with l. Anticipating this, 2 should choose L_2. Knowing that, player 1 will always play R_1.[13]

Whereas in the normal form condition (R) entails iterated elimination of dominated strategies, in the extensive form it constrains the possible interpretations of deviations. In particular, it requires beliefs to be consistent with sensible interpretations of a player's deviation from equilibrium, where a "sensible interpretation" is one that makes the deviation compatible with common knowledge of rationality. In figure 5, if player 2 gets to play, then player 1 must have forgone the payoff of 2 in favor of playing R_1. The only equilibrium in the subgame that yields a payoff greater than 2 to player 1 is (L_2l), hence 2 should deduce from the fact that node y is reached that 1 is planning to choose l at his next information set. If so, then 2's best reply is L_2, and player 1, anticipating player 2's reasoning, will conclude that it is optimal for him to play R_1.

What I have just described is a forward induction argument which, when coupled with condition (R), suggests that we interpret deviations as signals. For this interpretation to be consistent with rationality (and thus not to violate (R)), 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. Restricting deviations to undominated actions leads to the following iterated dominance requirement:

> (ID) A plausible equilibrium must remain plausible when a (weakly) dominated strategy is deleted.

Coupling the conditions (R), (BI), and (ID) merges the two seemingly different motivations behind the program for refining Nash equilibrium. The first motivation is to restrict out-of-equilibrium behavior, and hence to rule out deviations that do not have plausible explanations. The second motivation is to rule out equilibria that involve weakly dominated strategies and are therefore threat-vulnerable. The two motivations are only superficially different. If we think of restricting out-of-equilibrium beliefs, a very plausible restriction is to ask that beliefs be consistent with common knowledge of rationality. Common knowledge of rationality in turn implies that no player should ever be expected to choose a (weakly) dominated strategy. So equilibria that involve weakly dominated strategies should be ruled out.

REFINEMENTS

In the game of figure 5, I used a forward induction argument and interpreted player 1's choice as a signal to player 2. A question this argument raises is whether it is really so evident that there always exists a unique rational inference to draw from a player's off-equilibrium action. The same behavior, in

other words, could be explained in several ways, all of them compatible with a player being rational. A typical such case is that of non-cooperative games of imperfect information with multiple Nash equilibria. To identify a subset of "plausible" Nash equilibria, we have to check that a Nash equilibrium is robust to deviations. Even if we consider only those deviations that are consistent with common knowledge of rationality, there might be more than one way to make a deviation compatible with rational behavior. In this case further conditions should be imposed on out-of-equilibrium beliefs to obtain, whenever possible, a "ranking" of all the plausible explanations of deviations. To be able to eliminate all but one equilibrium and thus recommend a unique strategy for every player, game theorists must recommend a *uniquely rational* configuration of beliefs.[14] To do so, it is not enough to assume beliefs to be internally consistent. It must be further assumed that belief-rationality is a property resulting from the procedure by which beliefs are obtained, and it must be shown that there exists a rational procedure for obtaining them.

Game theorists have proposed various refinements of the Nash equilibrium concept to deal with this problem. Unfortunately, none of them succeeds in picking out a unique equilibrium across the whole spectrum of games.[15] Within the class of refinements of Nash equilibrium, two different approaches can be identified. One solution aims at imposing restrictions on players' beliefs by explicitly allowing for the possibility of error on the part of the players. This approach underlies both Selten's notion of 'perfect equilibrium',[16] and Myerson's notion of 'proper equilibrium'.[17] The alternative solution is based upon an examination of rational beliefs rather than mistakes. The idea is that players form conjectures about other players' choices, and that a conjecture should not be maintained in the face of evidence that refutes it. This approach underlies the notion of 'sequential equilibrium' proposed by Kreps and Wilson.[18] All of these solutions will be defined by means of examples next. For the moment, let us say that they all impose restrictions on players' beliefs, so as to obtain a unique rational recommendation as to what to believe about other players' behavior. This supposedly guarantees that rational players will select the equilibrium that is supported by these beliefs. Both approaches, however, fail to rule out some equilibria which are supported by beliefs that, although coherent, are intuitively implausible.

My objection concerns the nature of the restrictions imposed on players' beliefs. The specification of the equilibrium requires a description of what the agents expect to happen at each node, were it to be reached, even though in equilibrium play most of these nodes are never reached. The players are thus assumed to engage in counterfactual reasoning (from the viewpoint of the equilibrium under consideration) regarding behavior at each possible node.[19] For example, if in equilibrium a certain node would never be reached, a player asking himself what to do were that node to be reached is in fact asking himself why a deviation from that equilibrium would have

occurred. If in the face of a deviation he would still play his part in the equilibrium, then that equilibrium is "robust", or plausible. The game in figure 6 illustrates the reasoning process through which the players come to eliminate implausible (i.e., imperfect) equilibria.

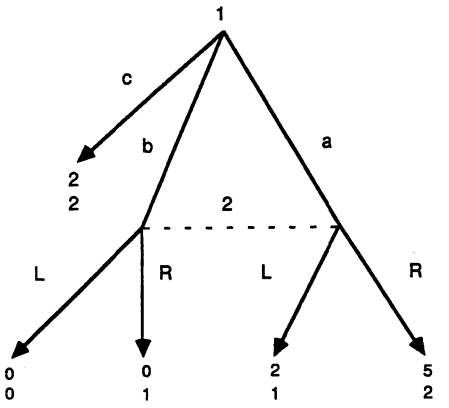

FIGURE 6

The game has two Nash equilibria in pure strategies, (c,L) and (a,R). Selten rejects equilibrium (c,L) as being unreasonable. To see how this conclusion is reached, let us follow the reasoning imputed to the players. In so doing, I expound Selten's well-known concept of *perfect equilibrium.*

Suppose that during preplay communication the players agree to play (c,L). Whether or not 1's choice of c is rational depends upon what he expects that 2 would do if he played a or b instead. For suppose that, contrary to 2's expectations, she is called to decide. Will she keep playing her equilibrium strategy? Evidently not, since L is strictly dominated by R. Thus, for any positive probability that a or b are played by 1, player 2 should minimize the probability of playing L. This reasoning will in fact take place even before the unexpected node is reached, since a rational player should be able to decide beforehand what it is rational to do at every possible node, including those which would occur with probability zero if a given equilibrium is played.

The players are reasoning counterfactually, asking themselves what they would do if a deviation from equilibrium were to occur, and understand that every information set can be reached, with at least a small probability, since it is always possible that a deviation from equilibrium play occurs by mistake. A sensible equilibrium will therefore prescribe rational (i.e., maximizing) behavior at every information set, since an equilibrium strategy must be optimal against some slight perturbations of the opponent's equilibrium strategies.[20] However, not all perfect equilibria are plausible, as the example in figure 7 illustrates.

There are two equilibria, (c,L) and (a,R), and they are both perfect. In particular, (c,L) is perfect if player 2 believes that player 1 will make mistake b with a higher probability than mistake a, but where both probabilities are very small, while the probability of 1 playing c will be close to one. If this is what 2 believes, then she should play L with probability close to one. But why should 2 believe that mistake b occurs with higher probability than mistake a? After all, both strategies a and c dominate b, so that there is little

reason to expect mistake b to occur more frequently than mistake a. Equilibrium (c,L) is perfect, but it is not supported by reasonable beliefs. The apparent limitation of the idea of perfectness is that restrictions are imposed only on equilibrium beliefs, while *out of equilibrium beliefs are unrestricted*: A player is supposed to ask whether it is reasonable to believe the opponent will play a given Nash equilibrium strategy, but not whether the beliefs supporting the other player's choice are rational.

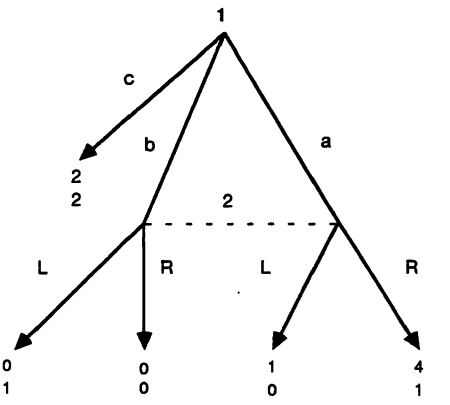

FIGURE 7

Let us compare for a moment the games of figures 6 and 7. In figure 6, the equilibrium (c,L) is ruled out because player 1 cannot possibly find any out-of-equilibrium belief supporting it. Player 2, facing a deviation, would never play the dominated strategy L. In Figure 7 instead, when player 1 wonders whether player 2 will keep playing L in face of his deviation, he can attribute a belief to player 2 that would justify her choice of L (in this game, 2 must believe that b has a greater probability than a). But player 1 does not ask whether the beliefs he attributes to player 2 about the greater or lesser likelihood of some deviation are at all justified. This, however, is a crucial question, since only by distinguishing those deviations (and out-of-equilibrium beliefs) that are more plausible from those that are less plausible is it possible to restrict the set of equilibria in a satisfactory way.

In order to restrict the set of equilibria, restrictions need to be imposed on all beliefs, including out-of-equilibrium ones. A player, that is, should only make conjectures about the opponents' behavior that are rationally justified, and he should believe that his opponents expect him to provide such a rational justification. It might be argued, for example, that a rational player will avoid costly mistakes. Thus a *proper* equilibrium need only be robust with respect to plausible deviations, meaning deviations that do not involve costly mistakes.[21] In the game of figure 7, if player 2 were to adopt this criterion she would assign deviation b a smaller probability than deviation a, and so she would play R with as high a probability as possible. This reasoning rules out equilibrium (c,L) as implausible.

An objection to this further refinement is the following: while this refinement rightly attempts to restrict out-of-equilibrium beliefs, it only partially succeeds in doing so. There are cases in which one mistake is more costly than another only insofar as the player who could make the mistake has definite beliefs about the opponent's reaction. As the game in figure 8 illustrates, these

beliefs require some justification, too.

Here both (a,R) and (c,L) 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 deviation b than to deviation a. If player 1 were to expect player 2 to behave in this way, mistake b would indeed be less costly than mistake a. In this case strategy L would be better for player 2. Thus b is less costly if 1 expects 2 to respond with L, and 2 will respond with L only if she can expect 1 to expect her to respond with L. But why should 2 be expected to play L in the first place? After all, strategy b is strictly dominated by c, which makes it extremely unlikely that deviation b will occur. So if a deviation were to occur, it would plausibly be a, and then player 2 would choose R. Hence the equilibrium (c,L) is unreasonable.

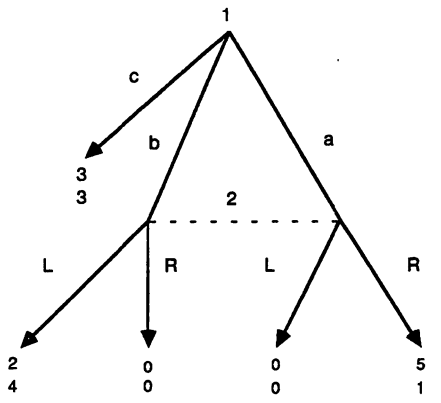

FIGURE 8

These examples suggest that for an equilibrium to be sensible, out-of-equilibrium beliefs need to be rationally justified. A player who asks himself what he would do in the face of a deviation must also find good reasons for that deviation to occur, which means explaining the deviation as the result of plausible beliefs on the part of both players. Hence a "theory of deviations" must rest upon an account of what counts as a plausible, or rational, belief.

Belief rationality, however, cannot reduce to coherence, or to the condition that a conjecture ought not to be maintained in the face of evidence that refutes it. These minimal rationality conditions are exploited by the *sequential equilibrium* notion,[22] which explicitly specifies beliefs at information sets lying off the equilibrium path. Briefly stated, a sequential equilibrium is a collection of belief-strategy pairs, one for each player, such that (i) each player has a belief (i.e., a subjective probability) over the nodes at each information set, and (ii) at any information set, given a player's belief there and given the other players' strategies, his strategy for the remainder of the game maximizes his expected payoff. More specifically, suppose that a given equilibrium is agreed upon and a deviation occurs. When a player finds herself at an unexpected node she will try to reconstruct what went wrong, but usually she will not be able to tell at which point of her information set she is. This uncertainty is represented by posterior probabilities on the nodes in her information set. When she acts so as to maximize her expected utility with respect to these beliefs, the player assumes that in the

rest of the game the original equilibrium is still being played. A sequential equilibrium has the property that if the players behave according to conditions (i) and (ii), no player has an incentive to deviate from the equilibrium at any information set. The problem with sequential equilibrium is that nothing is assumed about the plausibility of players' beliefs; that is, an equilibrium strategy must be optimal with respect to *some* beliefs, but not necessarily reasonable beliefs. So in the games in figures 6, 7, and 8, both Nash equilibria are sequential, since if player 1 chooses c, then any probability assessment by player 2 is reasonable. Such minimal rationality conditions are obviously too weak to rule out intuitively implausible beliefs.

A possible solution to the problem of eliminating implausible beliefs lies in combining the heuristic method implicit in the 'small mistakes' approach with the analysis of belief rationality characteristic of the sequential equilibrium notion. The 'small mistakes' approach features the role of anticipated actions off the equilibrium path in sustaining the equilibrium. In so doing, it models the players as engaged in counterfactual arguments which involve a revision of their original belief that a given equilibrium is being played.[23] For this process of belief change not to be arbitrary, it must satisfy some rationality conditions. Belief rationality should be a property of beliefs that are revised through a rational procedure. If there were a unique rational process of belief revision, then there would be a unique best theory of deviations that a rational player could be expected to adopt, and common knowledge of belief rationality would suffice to eliminate all equilibria which are robust only with respect to implausible deviations.

MODELING BELIEF CHANGES

In the foregoing examples, I eliminated implausible equilibria by checking each equilibrium's stability in the face of possible deviations. This method, which is common to all refinements of Nash equilibrium, is supposedly adopted by the players themselves before the start of the game, helping them to identify, whenever possible, a unique equilibrium. My counterexamples show not only that uniqueness is anything but guaranteed by those solutions, but also, and more important, that an answer to the problem of justifying equilibrium play is far from being attained. Indeed, as the games of figures 7 and 8 illustrate, players' expectations may be consistent, but they are hardly plausible. Perfect, proper, and sequential equilibria let players rationalize only some beliefs, in the absence of a general criterion of belief rationality that would significantly restrict the set of plausible beliefs. A criterion of belief rationality, it must be added, would have the twofold function of getting the players to identify a unique equilibrium as well as justifying equilibrium play. In what follows, I shall explicitly model the elimination of

implausible equilibria as a process of rational belief change on the part of the players.[24] In so doing, my aim is twofold. On the one hand, the proposed model of belief change has to be general enough to subsume the canonical refinements of Nash equilibrium as special cases. On the other hand, it must make explicit the conditions under which both the problem of justifying equilibrium play and that of attaining common knowledge of mutual beliefs can be solved.

The best-known model of belief change is Bayesian conditionalization: beliefs are represented by probability functions defined over sentences, and rational changes of beliefs are represented by conditionalization of probability functions. The process is defined thus: p' is the conditionalization of p on the sentence E if and only if, for every sentence H, $p'(H) = p(H\&E)/p(E)$. When $p(E) = 0$, the conditionalization is undefined. Since in our case a player who asks himself what he would do were a deviation to occur is revising previously accepted beliefs (e.g., the belief that a given equilibrium is played), and is accepting as new evidence a sentence E' whose prior probability is zero, conditionalization is not applicable in this context as a viable description of belief change.

Some have argued that conditionalization can be defined even if $p(E) = 0$, if one takes the conditional probability p(H/E) as primitive.[25] Nonetheless, it must be noted that conditionalization only applies to changes of beliefs where a new sentence is accepted that is not inconsistent with the initial corpus of knowledge, while the type of belief change I shall discuss involves a sentence E that is not a possibility, given the background knowledge of the players. The type of belief change I want to consider requires one to accept *less* than one did before in order to investigate some sentence that contradicts what was previously accepted. This type of belief change is "nonmonotonic", because when a new belief is added to one's stock of beliefs not all of one's previously accepted beliefs can be retained. Such changes are fairly common in hypothetical reasoning, and have been variously called "question opening"[26] and "belief contravening."[27] Gardenfors[28] has proposed a model of belief change that specifically focuses on the factors governing changes of beliefs when earlier accepted beliefs are retracted in order to add a new belief inconsistent with the previous belief system held by the agent. In what follows I will present a similar model to represent the virtual process of belief revision undergone by the players before playing the game.

As is customary in the refinements literature, equilibria are taken as reference points, and players are meant to believe that each of them can compute the Nash equilibria of the game. We may suppose that the players are told by the game theorist to play a particular equilibrium, or that instead, preplay communication is permitted and the players come to agree to play one equilibrium. In both cases, we ask whether the prescription or the agreement is reasonable, given that they are by definition self-fulfilling. Another way to put it is that we model the players as deciding by themselves (that is,

without the help of the game theorist or communication) whether an equilibrium is robust to deviations, repeating the procedure for every equilibrium of the game. At the end of the process, the players are left with a set (hopefully, a singleton) of stable equilibria that correspond to the set of reasonable, self-fulfilling agreements or prescriptions.

Let us assume that each player i starts this process of successive elimination of equilibria with a model of the game, denoted by M_i^0. This model is a *state of belief*, representable as a set of sentences expressed in a given language L. We assume L to be closed under the standard truth-functional connectives and to be governed by a logic *L* which contains all truth-functional tautologies and is closed under Modus Ponens. We call a subset M of L a *weakly rational belief set* if it satisfies the following conditions:

C1. M is non empty,
C2. if $A \in M$ and $B \in M$, then $A \& B \in M$,
C3. if $A \in M$ and $A \Rightarrow B$ is a truth-functional tautology, then $B \in M$.

C1–C3 imply that a weakly rational belief set M is a set of sentences from the language L which contains all logical truths of L and is closed under Modus Ponens. Such a set consists of all the sentences that an agent accepts in a given state of belief, where accepting a sentence means having full belief in it, or assigning to it probability one. The initial system of beliefs is a model of the game that includes axioms describing the rules of the game, players' strategies and payoffs, and the available solutions. The model might also include a statement to the effect that the players have agreed upon a particular solution.

To make this point clear, consider again figure 7 and suppose that the initial model of the game M_i^0 ($i = 1, 2$) includes a statement to the effect that the equilibrium (c,L) is being played.[29] The model will also contain the following set of sentences:

(i) the players are rational (e.g., expected utility maximizers);
(ii) the players always play what they choose;
(iii) player 1 chooses to play c;
(iv) player 1 plays c;

For all i, we assume M_i^0 to be common knowledge.

To decide whether the equilibrium being considered is plausible, a player will ask himself what the other would do if he were to reach an unexpected information set, that is, an information set that would never be reached if the equilibrium (c,L) were played. In order to consider the possibility of a deviation, the player has to eliminate from M_i^0 some beliefs, so that his beliefs no longer entail the impossibility of that deviation. The player will thus have to *contract* his original belief set by giving up his belief in sentence (iv), but since he has to comply with the requirement that a belief set be

closed under logical consequence, he may have to relinquish beliefs in other sentences as well.

In general there will be many ways to fulfill this requirement. For example, since (iv) is implied by the conjunction of (ii) and (iii), eliminating (iv) implies eliminating the conjunction of (ii) and (iii). This means eliminating (ii), or eliminating (iii), or eliminating both. Besides maintaining consistency, it seems reasonable to require belief changes to satisfy a further rationality criterion: that of avoiding unnecessary loss of information. Assuming this further requirement, the players face two minimal choices compatible with the elimination of (iv): either (iv) and (iii) are eliminated, or (iv) and (ii).

A criterion of informational economy can be interpreted in several ways. If we think of information as an objective notion, the information contained in a corpus of knowledge is a characteristic of that corpus independent of the values and goals of the agents, whereas informational value is the utility of the information contained. That one piece of information is more useful than another does not mean that it is better confirmed, more probable, or even more plausible. In fact, we want to distinguish between *degrees of acceptance* and *degrees of epistemic importance*.[30] If we define M^i as a set of sentences whose falsity agent i is committed to discount as a serious possibility, all the sentences in M^i will have the same degree of acceptance, in the sense that all will be considered maximally probable, but their degrees of epistemic importance (or epistemic utility) will differ according to how important a sentence is to inquiry and deliberation. For example, if explanatory power is an important element in an agent's decision framework, then a lawlike sentence will be epistemically more important than an accidental generalization, even if their relative importance cannot be measured in terms of truth values, since the agent will be equally committed to both insofar as they are part of his belief system.

When M_i^0 is contracted with respect to some beliefs, we obtain a new belief set M_i^1 which contains less information than the original belief set. The objective notion of information allows partial ordering of belief sets with respect to set inclusion: if M is a proper subset of M', the information contained in M' is greater than the information contained in M. Minimum loss of information in this sense means eliminating as little as possible while maintaining consistency. Considering the utility of information means eliminating first all those sentences that have lower informational value. Note that using a criterion of informational value may or may not complete the partial ordering with respect to information: whenever M is a proper subset of M', the informational value carried by M' cannot be less than that carried by M, but it may be the same.

Whatever interpretation is attributed to the criterion of informational economy, every contraction of a belief set will have to satisfy a minimal set

of further (weak) rationality conditions. Let us denote a *contraction* of a belief set M with respect to a sentence A by M_{-A}. Every contraction M_{-A} must satisfy the following conditions:

C4. M_{-A} is a belief set,
C5. $M \supseteq M_{-A}$,
C6. $A \notin M_{-A}$ unless A is logically valid,
C7. if $A \notin M$, then $M_{-A} = M$,
C8. if A and B are logically equivalent, $M_{-A} = M_{-B}$.

An *expansion* of a belief set M with respect to a sentence A, denoted by M_{+A}, similarly must satisfy:

C9. M_{+A} is a belief set,
C10. $M_{+A} \supseteq M$,
C11. if $A \in M$, then $M_{+A} = M$,
C12. if A and B are logically equivalent, $M_{+A} = M_{+B}$.
C13. if $A \in M$, then $(M_{-A})_{+A} \supseteq M$.

The changes of beliefs I am discussing involve accepting a sentence the negation of which was earlier accepted; such belief contravening changes can be analyzed as a sequence of contractions and expansions. Suppose $\sim A \in M$. Then in order to add a belief contravening statement A, one will first contract M with respect to $\sim A$, and then expand $M_{-\sim A}$ by A. We define M_A as $(M_{-\sim A})_{+A}$. Let us call the revised belief set M_A a *counterfactual change* of M. Indeed, when a player asks "If there were a deviation from the equilibrium strategy c, then . . ." she is asking a counterfactual question (from the viewpoint of the model of the game she starts with). Answering the counterfactual means first contracting and then expanding the original model of the game. A basic acceptability criterion for a sentence of the form, "If A were the case, then B would be the case" is that this sentence is acceptable in relation to a state of belief M if and only if B is accepted in a revised belief set M_A which results from minimally changing M to include A (i.e., iff $B \in M_A$).[31]

It remains to be established how a revised belief set is to be constructed. Since we want the contraction of the belief set M with respect to $\sim A$ to be minimal, in order to lose as little information as possible, we want $M_{-\sim A}$ to be as large a subset of M as possible. Let us define $M_{-\sim A}$ as *maximally consistent* with A in relation to M iff for every $B \in M$ and $B \notin M_{-\sim A}$, $\sim (A \& B) \in M_{-\sim A}$. So if $M_{-\sim A}$ were expanded by B, it would entail $\sim A$. Still there might be many subsets of M which are maximally consistent with A. As the above interpretation of what a minimal change of beliefs consists of is usually not strong enough to isolate a unique answer to the counterfactual question, the rationality conditions we have imposed thus far on belief sets and contractions of belief sets do not guarantee that the players will revise

their beliefs in the same way, thus ending up with the same model of play and the same solution set.

All the contracted sets thus obtained will be proper subsets of the original belief set, but the ordering of set inclusion will in general be partial. Wanting the ordering to be complete is a good reason to introduce further restrictions. Another reason for supplementing the criterion of maximal consistency is the following. Suppose that statement A is contained in a corpus of knowledge M and that there is a statement B which has nothing to do with A. Then M will also contain both disjunctions $A \vee B$ and $A \vee \sim B$. If M is minimally contracted with respect to A, then either $A \vee B$ or $A \vee \sim B$ will belong to M_{-A}. If M_{-A} is expanded by $\sim A$, $(M_{-A})_{+\sim A}$ will contain either B or $\sim B$. Hence revised belief sets obtained from maximally consistent contractions will contain too much, since for every sentence in L, either it or its negation will be in the revised belief set. [32]

Since different contraction strategies will differ from one another with respect to the loss of informational value incurred, it seems reasonable to supplement maximal consistency with a criterion of minimum loss of informational value. In this case, it must be specified how sentences can be ordered according to their informational value or epistemic utility. If we assume that all the sentences in an agent's belief set have the same degree of acceptance, it will be impossible to set them apart in terms of probability, evidential support, or plausibility. When judging the loss of informational value engendered by a contraction, what is at issue is not the truth value of the different items, but their relative importance with respect to the objectives of the decision maker.

I must emphasize that the idea of an ordering based upon informational value has both advantages and drawbacks. For one, informational value can be an extremely subjective criterion, depending as it does on the context (for example, the type of game being played), the individual's understanding of the situation he is in (players may have different information about the game), and one's aims in a given context. To be at all helpful in selecting an equilibrium, an ordering of informational value must be shared by the players, and this fact must be common knowledge. On the other hand, if players have homogeneous ends and a common understanding of the situation they are in, an ordering according to informational value seems preferable to others. For example, it is generally assumed that players are endowed with a common theory of the game and common knowledge of rationality and payoffs. Since one's choice depends on what one expects the opponents to do, one's goal is that of predicting as best as possible the opponents' play. And because rationality is common knowledge, a deviation from equilibrium play will generally be explained by either dropping the assumption that players play what they choose, or by assuming that the deviation was intentional. This latter assumption has the advantage of letting players predict the future moves of the intentional deviator, something that is not easy to do

when deviations are interpreted as mistakes. In a game-theoretic context, to allow for mistakes when alternative explanations by intentional action are available entails a loss of predictive power. An ordering according to informational value thus provides a way of plausibly ordering refinements according to how much of the predictive power of the original theory of the game they give away.

On my interpretation, informational value is a pragmatic concept. Depending on the context, certain statements will be less vulnerable to removal than others, and in any context (e.g., in any belief state) it is possible to order the statements with respect to their epistemic entrenchment (or epistemic importance). In a given context, epistemic entrenchment can be characterized by a complete preorder (a reflexive and transitive relation) over sentences which tells which sentences are more epistemically entrenched than others. This ordering influences belief revisions in that any revision should retain more entrenched beliefs in preference to less entrenched ones.

We write $a \leq b$ to mean that sentence b is at least as epistemically entrenched as a. Though epistemic entrenchment cannot be measured in a quantitative way, Gardenfors[33] has provided the following postulates for its qualitative properties:

($\leq$ 1) If $a \leq b$ and $b \leq c$, then $a \leq c$; (transitivity)
($\leq$ 2) If $a \vdash b$, then $a \leq b$;[34] (dominance)
($\leq$ 3) Either $a \leq a \wedge b$ or $b \leq a \wedge b$; (conjunctiveness)
($\leq$ 4) If M is consistent, then $a \leq b$ for all b iff $a \notin M$; (minimality)
($\leq$ 5) If $a \leq b$ for all a, then $\vdash b$. (maximality)

Axiom ($\leq$ 1) says that $\leq$ is a transitive relation. Axiom ($\leq$ 2) says that if a entails b, then retracting a is a smaller change than retracting b, since the closure requirement on belief sets means that b cannot be retracted without retracting a as well. Axiom ($\leq$ 3) says that a conjunction cannot be retracted without giving up at least one of its conjuncts. Taken together the first three axioms imply that $\leq$ is a connected ordering. According to ($\leq$ 4), sentences which are not in a belief set are minimally entrenched in that set, and according to ($\leq$ 5) a proposition is maximally entrenched only if it is logically valid.

There are two conditions that relate entrenchment orderings to contraction functions:

(α) $a \leq b$ iff either $a \notin M_{-a\&b}$ or $\vdash a \,\&\, b$;
(β) $b \in M_{-a}$ iff $b \in M$, and either $a < a \vee b$ or $\vdash a$.

Condition (α) says that in contracting a belief set with respect to a conjunction we must give up the least epistemically entrenched conjunct; if both conjuncts are equally entrenched, then both should be retracted. Condition (β) characterizes contraction functions in terms of orderings of epistemic entrenchment.

Note that being able to order sentences by epistemic importance does not give an ordering of sets of sentences. If a set of sentences is finite, we can identify the informational value of the set with the informational value of the sentence which is the conjunction of all the sentences in the set. Our contracted belief sets, however, are not finite, in that they contain all logical truths. But since they generally are finitely axiomatizable, we can still rank belief sets in terms of the epistemic entrenchment of the conjunction of their axioms.

Let us now state the rules according to which a rational player will modify her beliefs:

R1. *any revised belief set satisfies C1–C13,*
R2. *from the set* $M_{-\sim A}$ *of all maximally consistent contractions of* M *with respect to* $\sim$A, *select the subset* $M^*_{-\sim A}$ *of the most epistemically important belief sets with the aid of the criterion of minimum loss of informational value,*
R3. *the new contracted belief set* $M_{-\sim A}$ *will include all the sentences which are common to the elements of* $M^*_{-\sim A}$, i.e., $M_{-\sim A} = \cap M^*_{-\sim A}$,[35, 36]
R4. *expand the belief set* $M_{-\sim A}$ *thus obtained by adding* A.

Note that R1 corresponds to the weak rationality criteria imposed on belief sets whereas R2 involves a stronger, substantive rationality criterion. It implies, for example, that it is always possible to define an ordering of epistemic entrenchment, however pragmatic and context dependent it may be. Furthermore, in any given game the ordering of sentences with respect to epistemic entrenchment must be unique, otherwise the players may never get to agree on the same interpretation of a deviation from equilibrium play. This requirement is not too far-fetched, since in real-life situations the context in which the interaction takes place often suggests a definite interpretation for out-of-equilibrium actions, hence for the reasons that underlie them. R2 says that a criterion of epistemic entrenchment may not avoid ties, in that there might be several belief sets that are 'most important' in this sense. If there are ties, R3 says that the contracted belief set should include all the sentences that are common to the "most epistemically important" belief sets. I assume these rules to be common knowledge among the players.

SIGNALS OR MISTAKES?

Consider again figure 7. Before playing the game, the players will consider in turn each Nash equilibrium of the game as a candidate for a solution and, for each such candidate, ask whether it is robust to deviations. As I already mentioned, this process of elimination of "weak" equilibria is crucial to a player's choice of a strategy. Consider player 2. If she assumes that (c,L) is

being played, a question she will ask is whether it makes sense to keep playing L once a deviation has occurred. Her answer (and her strategy choice) will depend on how she explains the deviation. The model of belief revision I have just described is a way to explicitly model the process of elimination of "weak" equilibria that leads to the choice of a strategy. Let us assume that, besides the usual common knowledge assumptions, it is common knowledge among the players that each player undergoes a process of "testing" the different Nash equilibria and that both players adopt the same model for belief revision. Moreover, the players have a shared ordering of epistemic entrenchment, and this fact is common knowledge. It follows that each player can replicate the reasoning (and the conclusions) of the other.

Consider, again, player 2's reasoning (in figure 7). If she considers first the equilibrium (c,L), her initial model M_2^1 will include the following set of sentences:

(i) the players are rational (e.g., expected utility maximizers);
(ii) the players always play what they choose;
(iii) player 1 chooses to play c;
(iv) player 1 plays c;

Assuming that a deviation takes place, player 2 must decide how to contract her original model M_2^1 with respect to sentence (iv) in order to retain consistency. If (iv) is retracted according to R2, she is left with two maximally consistent belief sets: M' = (i), (iii) and M"= (i), (ii). In order to complete the ordering, she has to assess which of the two contractions entails a greater loss of informational value or, if there is a tie, she proceeds to apply R3. The last step consists in adding to the belief set thus obtained the negation of sentence (iv). Player 2 will then choose that strategy that is a best reply to her revised belief set.

M' entails substantial informational loss, since eliminating (ii) introduces an ad hoc element in the explanation of behavior. Retaining the assumptions that player 1 is rational and chooses to play the equilibrium strategy c means explaining a deviation as the effect of a random mistake (indeed, systematic mistakes would be incompatible with rationality). Hence even if player 1 were to make a long series of mistakes, they would be interpreted as random and uncorrelated, and each one would have to be separately explained. But then any arbitrary pattern of play is compatible with rational behavior, with the consequence of undermining the predictive strength of a principle of rationality. Moreover, taking mistakes seriously as clues to a basic flaw in a player's rationality or judgment might prove advantageous. In chess, for example, bad play is unlikely to be due to transient muddled judgment or to a persistent, random muscle twitch that impedes a player from making the right move. What is likely is that a mistake increases the odds of subsequent mistakes since it is a symptom of poor judgment, and hence a symptom of decisions which are less than optimal. Thinking that the next move of such

a player will be optimal may not just be a misjudgment: it may make one lose the opportunity to exploit this pattern of play.

A further consideration weakens the plausibility of M'. In the game in figure 7, choosing this contraction (and then expanding the belief set) implies maintaining that both equilibria (a,R) and (c,L) are sensible. But (c,L) is plausible only if player 1 expects player 2 to play L and player 2, in turn, plays L only if she expects deviation b to occur more frequently than deviation a. In order to play c, player 1 must then believe that 2 thinks deviation b to be more likely than deviation a. If player 2 thinks that deviation b is more likely than deviation a, she must either believe that player 1 is irrational (since strategy b is dominated by a and c), or she must believe that player 1 believes that she believes he is irrational. Since the contracted belief set M' retains the belief that 1 is rational and that he chooses strategy c, 2 must assume that 1 plays c because he believes that 2 believes he is irrational. Nothing in 2's original belief set justifies this belief. M' thus involves arbitrary beliefs, since it does not require out-of-equilibrium beliefs to be plausible.

That consistently bad play is unlikely to be explained by muscle twitches is prima facie an argument in favor of dropping the assumption that players are always rational. Before giving up a rationality assumption, however, it is reasonable to consider alternative explanations that make deviations compatible with rational play. M" provides such an explanation. By eliminating (iii), M" allows player 2 to interpret deviations as actions intended to influence other players' beliefs (let us call such actions "signals"). Rational player 1 will only deviate if he expects to get a higher payoff than what he gets by playing c. And he expects a higher payoff only if he is reasonably sure that player 2 interprets his move as a rational choice. The revised belief set obtained from M" entails that, facing a deviation, player 2 will interpret it as a signal from player 1: upon reaching her information set player 2 infers that player 1 believes that she will respond with R.

Player 2 must now choose between two belief sets, M" and M'. She must decide which of the two belief sets is the most epistemically important by weighing the loss of informational value due to removing sentence (iii) against the loss involved in eliminating (ii). As I pointed out, it is more costly (in terms of epistemic utility) to assume that players make mistakes, since this assumption weakens the predictive (and explanatory) power of the rationality assumption. Player 2 will therefore choose the contracted belief set M". Since the rules for belief revision and the ordering of epistemic importance are common knowledge among the players, both players will exclude the (c,L) equilibrium and choose (a,R).

This type of belief revision corresponds to the forward induction argument put forth by Kohlberg and Mertens[37] according to which a deviation should be interpreted, whenever possible, as the intentional move of a rational player. What the belief revision model presented here adds is a

theoretical justification, based on a definition of substantive belief rationality, for preferring a forward induction argument to those that focus on mistakes or rely on arbitrary probability assignments. The belief revision model implies that the revised belief set obtained from contraction M" gets the players to play the equilibrium (a,R) in the games of figures 6, 7, and 8, since in each of them player 1 has a good reason for deviating from his equilibrium strategy c. This means, from the game theorist's viewpoint, that in all three games there exists a unique rational recommendation of beliefs regarding the other player's behavior.

Is it possible, just by assuming players to have common knowledge of the rules R1-R4, to predict that they will agree on the same revised belief set? The answer is affirmative only when the ordering of sentences with respect to epistemic entrenchment is uncontroversial. As I pointed out, this ordering is not connected with the probabilities of the sentences in a belief state, since in any given belief state all sentences are maximally probable and they only differ in their explanatory power and informational value. The assumption that players are expected utility maximizers, for example, is difficult to give up since it is a powerful explanatory (and predictive) principle. Of the two contractions examined here, M" involves the least loss of informational value, in that it not only avoids ad hoc explanations, but it also extends rationality to out-of-equilibrium behavior, avoiding the implausible beliefs entailed by M'. In this case the interpretation of minimum informational loss is straightforward, since a statement about a regularity of behavior (i.e., "the players always play what they choose") is more epistemically entrenched than a statement of less general scope (i.e., "player 1 chooses to play strategy c").

There is a correspondence between the rules for rational belief revision and the principles I have put forth for eliminating implausible equilibria. In the former examples, the equilibrium identified by the criterion of minimum loss of informational value is the same as that which results from iterated elimination of (weakly) dominated strategies. Note that in all three games condition (BI) does not apply, since there are no proper subgames; hence if strategy c is expected, L remains a best reply in all three cases, because if c is played it does not really matter what player 2 does. Condition (R) is relevant since it rules out dominated strategies iteratively, and condition (ID) lets the players choose that equilibrium which survives iterated elimination of (weakly) dominated strategies. Both conditions conform to an interpretation of deviations as the result of intentional moves on the part of rational players, hence to a way of changing one's beliefs that preserves the ones that carry the greatest informational value.

Even if I succeeded in establishing on firmer grounds the plausibility of forward induction, I want to stress that it cannot be concluded that this particular model of belief revision (or, for that matter, any other model) *implies* forward induction. The particular refinement one may wish to apply will always depend upon one's hypotheses about out-of-equilibrium actions, and

such conjectures in turn depend upon the context of play, which is not captured by the formal model. Moreover, a model of the game that includes a procedure for belief revision assumes, in addition to common knowledge of rationality and of the structure of the game, common knowledge of the rules for belief revision and of a shared order of epistemic entrenchment. There are circumstances in which such strong common knowledge assumptions may be unrealistic. An unexpected move by a clever opponent is easily interpreted as a signal, but what about the same move made by someone that comes from a different culture? External considerations will lend plausibility to the selection of sentences we include in the model of the game, as well as to the arguments favoring some given ordering of informational value, but obviously these will hardly be the only reasonable possibilities.

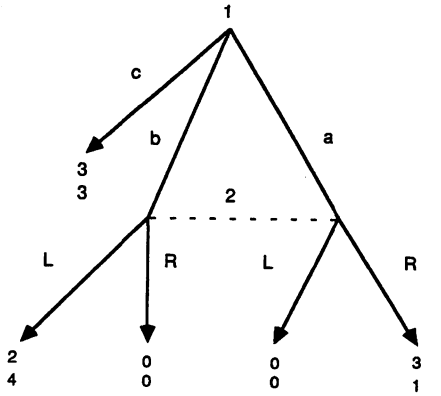

FIGURE 9

We may well ask how the theory of belief revision proposed here applies to those cases in which there is no good explanation for out-of-equilibrium behavior, i.e., when a deviation from equilibrium play is not explainable as an intentional choice. Consider the game in figure 9.

Suppose player 2 considers first equilibrium (c,L) as a candidate for a solution and asks what would happen were a deviation to occur. How will she contract her original belief set? The minimal revisions still involve deleting either sentences (iv) (player 1 plays c) and (iii) (he chooses to play c), or (iv) and (ii) (players always play what they choose). Now, however, eliminating (iii) and (iv), and then expanding by ~(iv), implies that player 1 chose not to play c. Since the best outcome player 1 can get by deviating is a payoff of 3, which is what he could get with certainty by playing strategy c, an intentional deviation seems unreasonable. Alternatively, one might argue that player 1 is really indifferent between c and a, and lets his choice be determined by the toss of a coin. But this implies that he is quite sure that player 2 has eliminated the possibility of his playing b (which is dominated) and sees the game as he does. Even if they share the same rules for belief revision, it is not obvious that a rational player should be indifferent between c and a, so player 1's faith in player 2's replying with R cannot be grounded in a shared view of what it is rational to do in such circumstances. It seems that player 1, if rational, should not be so sure of player 2's interpretation of his deviation, in which case it is safer for him to play c.

Eliminating (iv) and (iii), and then expanding by ~(iv), entails a greater loss of informational value than the alternative revision strategy, which eliminates (iv) and (ii), as it makes an intentional deviation from c incompatible with assuming that player 1 is rational. The only alternative contraction involves deleting (iv) and (ii), so that the revised belief set obtained by adding ~(iv) corresponds precisely to the 'small mistakes' hypothesis. At this point, the choice of player 2 depends on the probability she assigns to each deviation. If 2 believes mistake a will have a chance of occurring greater than 4/5, she will play R, and she will play L if the chance of b occurring is greater than 1/5. Since the theory of belief revision shared by the players entails that a deviation would only occur by mistake, all the conjectures held by player 2 are equally plausible. In this case there is no way to distinguish between the three equilibria (c,L), (c,R), and (a,R), which are in fact all perfect, proper, and sequential. The 'small mistakes' hypothesis is therefore a special case of our belief revision model, and it becomes plausible when deviations from equilibrium play cannot be interpreted as signals.

It might still be argued that if a player is perfectly rational she will not be prone to mistakes, be they temporary carelessness or confusion about what one is expected to do.[38] If the game is one of perfect information, a rational player should never be expected to deviate by mistake, or because she has a mistaken theory about the play of the game. If we follow this line of thought, we may think unreasonable all those equilibria that depend upon players' inability to interpret deviations as signals. Moreover, we may assume that a deviation for which no sensible interpretation exists never occurs. But if there are no plausible out-of-equilibrium beliefs, then behavior off the equilibrium path should not be thought to be possible. So for example the equilibrium (c,L) in figure 9, though self-enforcing, cannot be supported by plausible out-of-equilibrium beliefs.

CONCLUSIONS

Until now, I have assumed that the players have the same model for belief revision and share the same ordering of epistemic entrenchment. These assumptions are obviously quite strong. Since any ordering of epistemic importance is contextual, the players must have a similar understanding of the situation they are in, and this may be too much to ask in a new environment or when the players know very little about each other. In addition, players' beliefs are understood to be rational in the double sense of being consistent and of fulfilling several substantive criteria of rationality like those embedded in the rules for belief revision. What I am after, however, is not realism of the assumptions. I want, rather, to specify the logical conditions under which a Nash equilibrium can be inferred from assumptions

of common knowledge of the structure of the game and of players' mutual rationality. We know that in all those games that cannot be solved by dominance (or iterated dominance), common knowledge of the structure of the game and of players' mutual rationality are not sufficient to identify one Nash equilibrium as the "obvious solution" to the game.[39] To identify (and predict) one Nash equilibrium as the solution to the game, players' beliefs have to be specified. Short of assuming that beliefs are correct a priori, one has to spell out the process through which they become correct (and common knowledge), hence one has to extend the concept of belief rationality to include a model of how the players revise their beliefs out of equilibrium. If the process of belief revision identifies a unique solution to the game, then players who revise their beliefs according to that process should be expected to agree on that solution.

We may think of a model for belief revision as part of a complete theory of the game. Such a theory must explain, among other things, why the unexpected may happen. And if there are reasons why unexpected actions might be chosen, then these reasons should be part of the theory. A complete theory of the game should then assume that players are both practically rational (e.g., they are expected utility maximizers) and epistemically rational (e.g., their beliefs are the outcome of a rational process of belief formation). The theory of the game proposed here succeeds in identifying a unique equilibrium as "the most plausible solution" in many games in which other criteria for equilibrium selection (such as perfectness or properness) leave a set of equilibria.

NOTES

1. I would like to thank Sergiu Hart, Motty Perry, Shmuel Zamir, and especially Jim Joyce and Bart Lipman for many useful comments. This article appears in my book *Rationality and Coordination* (Cambridge: Cambridge University Press, 1993).
2. One could argue that the real key is common knowledge of actions, not of beliefs. If I know your beliefs, I don't necessarily know what action you will choose since you may be indifferent between two or more actions. Here we focus on games where the players never move simultaneously, so this indifference issue never really arises.
3. C. Bicchieri, "The Epistemic Foundations of Nash Equilibrium," in D. Little, ed., *On the Reliability of Economic Models: Essays in the Philosophy of Economics* (Kluwer, 1993).
4. The relation $<$ is asymmetric, transitive, and satisfies the following property: if $t < t''$ and $t' < t''$ and $t \neq t'$, then either $t < t'$ or $t' < t$. These assumptions imply that the precedence relation is only a partial order, in that two nodes may not be comparable, and that each node (except the initial node) has just one immediate predecessor, so that each node is a complete description of the path preceding it. When a node is not a predecessor of any node we call it a "terminal node".
5. If t and t' belong to the same information set, we require that the same player moves at t and t'. Also, a player must have the same set of choices at each node belonging to the same information set.

6. A strategy is weakly dominant if it gives a player payoffs that are greater or equal to the payoffs of any other strategy.
7. Note that in games in which a player has to move at least twice, one of them chronologically after the other, a strategy has to specify actions even after histories which are inconsistent with that very strategy.
8. C. Bicchieri, "Knowledge-Dependent Games: Backward Induction," in C. Bicchieri and M. L. Dalla Chiara, eds., *Knowledge, Belief, and Strategic Interaction* (Cambridge: Cambridge University Press, 1992).
9. A subgame is a collection of branches of a game such that they start from the same node and the branches and the node together form a game tree by itself.
10. Under act/state independence, rationality as admissibility is entailed by rationality as expected utility maximization: a strictly dominated action is not a best reply to any possible subjective assessment, therefore an expected utility maximizer will never choose it.
11. Kohlberg and Mertens characterize a forward induction argument as follows: "A subgame should not be treated as a separate game, because it was preceded by a very specific form of preplay communication—the play leading up to the subgame" (E. Kohlberg and J. F. Mertens, "On the Strategic Stability of Equilibria," *Econometrica* 54 [1986]: 1013).
12. R. Selten, "Spieltheorethische Behandlung eines Oligopolmodels mit Nachfragetragheit," *Zietschrift fur die gesante Staatwissenschaft* 121 (1965): 301–24.
13. For $(R_1 | L_2)$ to obtain, common knowledge of rationality is not even needed. It is sufficient that player 2 knows that 1 knows a) that player 2 is rational, and b) that player 2 knows that 1 is rational.
14. Note that a family of permissible belief states would also do the job, provided its elements all determine the same equilibrium choice.
15. E. van Damme, *Refinements of the Nash Equilibrium Concept* (Springer Verlag, 1983) and E. van Damme, *Stability and Perfection of Nash Equilibrium* (Springer Verlag, 1987).
16. R. Selten, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 4 (1975): 22–55.
17. R. Myerson, "Refinements of the Nash Equilibrium Concept," *International Journal of Game Theory* 7 (1978): 73–80.
18. D. Kreps and R. Wilson, "Sequential Equilibria," *Econometrica* 50 (1982): 863–94.
19. H. S. Shin, "Counterfactuals, Common Knowledge, and Equilibrium," mimeo., Nuffield College, Oxford (1987); C. Bicchieri, "Strategic Behavior and Counterfactuals," *Synthese* 76 (1988): 135–69.
20. More precisely, a perfect equilibrium can be obtained as a limit point of a sequence of equilibria of disturbed games in which the mistake probabilities go to zero. Thus each player's equilibrium strategy is optimal both against the equilibrium strategies of his opponents and some slight perturbations of these strategies. See R. Selten, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games."
21. Myerson, op. cit.
22. D. Kreps and R. Wilson, op. cit.
23. R. Selten and U. Leopold have explicitly discussed the role of counterfactual reasoning in decision theory and game theory. See their "Subjunctive Conditionals in Decision and Game Theory," in W. Stegmuller, W. Balzer, and W. Spohn, eds., *Philosophy of Economics* (Berlin: Springer Verlag, 1982). Their model is a variant of the Stalnaker-Lewis theory of counterfactuals, which identifies the proposition expressed by a counterfactual conditional with a set of possible worlds and provides a selection function that selects the most similar world in which the conditional is true. See R. C. Stalnaker, "A Theory of Conditionals," in N. Rescher, ed., *Studies in Logical Theory* (Oxford: Blackwell, 1968) and D. Lewis, *Counterfactuals* (Cambridge, Mass.: Harvard University Press, 1973). Since the function selects among the possible worlds that make the antecedent of the conditional true, the one which is "closest" or "most similar" to the

actual world, it presupposes an ordering of possible worlds in terms of similarity with the actual world. The difficulty with this theory lies in the arbitrariness of the notion of similarity among worlds.

24. C. Bicchieri, "Strategic Behavior and Counterfactuals"; C. Bicchieri, "Self-refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis* 30 (1989): 69–85.
25. This argument originates with de Finetti. For an extensive defense of this idea, see Isaac Levi, "Coherence, Regularity and Conditional Probability," *Theory and Decision* 9 (1978): 1–15.
26. W. Harper, "Rational Conceptual Change," *PSA* (1977), 1976, vol. 2, *Philosophy of Science Association,* 462–94.
27. N. Rescher, *Hypothetical Reasoning* (North Holland, Amsterdam, 1964); I. Levi, "Subjunctives, Dispositions and Chances," *Synthese* 34 (1977): 423–55.
28. P. Gardenfors, "Conditionals and Changes of Belief," *Acta Philosophica Fennica* 30 (1978): 381–404; P. Gardenfors, "Rules for Rational Changes of Belief," *Philosophical Essays Dedicated to Lennart Aqvist on His Fiftieth Birthday,* Department of Philosophy, University of Uppsala (1982), 34.
29. It does not matter which equilibrium a player starts by considering, since she will have to repeat the reasoning for each equilibrium.
30. I. Levi, "Subjunctives, Dispositions and Chances"; I. Levi, "Serious Possibility," *Essays in Honor of Jaakko Hintikka* (Dordrecht: Reidel, 1979), 219–36.
31. This is a formulation of the Ramsey test that does not involve including conditional sentences as elements of belief sets. The idea that beliefs in conditional sentences lack truth value is advocated by I. Levi, and has the advantage of making the Ramsey test consistent with the *preservation criterion*, which says that if A is accepted in the belief set M and B is consistent with M, then A must be accepted in the minimal change of M needed to accept B. See I. Levi, "Subjunctives, Dispositions and Chances", and Levi, *The Enterprise of Knowledge* (Cambridge, Mass.: MIT Press, 1980). For an extensive discussion of these topics, see P. Gardenfors, "Conditionals and Changes of Belief," and Gardenfors, "Belief Revisions and the Ramsey Test for Conditionals," *Philosophical Review* 95 (1986): 81–93.
32. This difficulty is also pointed out in Gardenfors, "Epistemic Importance and Minimal Changes of Belief," *Australasian Journal of Philosophy* 62 (1984): 136–57, and in C. E. Alchourron, P. Gardenfors, and D. Makinson, "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions," *The Journal of Symbolic Logic* 2 (1985): 510–30.
33. P. Gardenfors, *Knowledge in Flux* (Cambridge, Mass.: MIT Press, 1988).
34. $a \vdash b$ means that there is a proof of b from a in language L.
35. The properties of this type of contraction function are discussed in Alchourron, Gardenfors, and Makinson, op. cit.
36. This intersection is not empty because all tautologies will be in the intersection, as well as the rules of the game and, in general, all the statements that belong to the players' background knowledge.
37. E. Kohlberg and J. F. Mertens, op. cit.
38. J. Hillas, "Sequential Equilibria and Stable Sets of Beliefs," mimeo., Stanford University (1987).
39. C. Bicchieri, "The Epistemic Foundations of Nash Equilibrium."