COUNTERFACTUALS AND BACKWARD INDUCTION*

Cristina Bicchieri

Introduction

In this paper I want to explore whether the assumption that players' rationality is common knowledge among them leads to inconsistencies in a special class of games. Simply stated, for a group of individuals to have common knowledge that p means that everybody knows that p is true, and everybody knows that everybody knows it, and so on ad infinitum. Common knowledge assumptions are standard in game theory. In normal form games, for example, the players are endowed with common knowledge of the rules of the game, and of their respective preferences and beliefs. Beliefs are about exogenous uncertainty, as well as endogenous uncertainty about the other players' choices and beliefs, which include beliefs about each other's rationality. If we exclude the case in which one or more players have dominated strategies, hence no common knowledge of beliefs needs to be assumed, in general every Nash equilibrium is supported by a configuration of beliefs which are common knowledge among the players.[1] Are common knowledge assumptions also needed in extensive form games? In the present paper, I only consider finite, extensive form games of perfect information. In such games, the classical equilibrium solution is obtained by backward induction. The solution is unique, and it is derived from a set of assumptions about the players' rationality and their mutual beliefs about each other's rationality. These assumptions, together with a specification of the structure of the game, and the hypothesis that the structure of the game is common knowledge, constitute the 'theory' of the game. It has been argued that if the rationality assumption is made common knowledge, the theory of the game will become inconsistent at some information set [Reny: 1987]. I have shown elsewhere that common knowledge of beliefs (and therefore of rationality) is neither necessary nor sufficient to obtain the backward induction solution [Bicchieri: 1989]. In fact, only distributed or full knowledge of the theory's

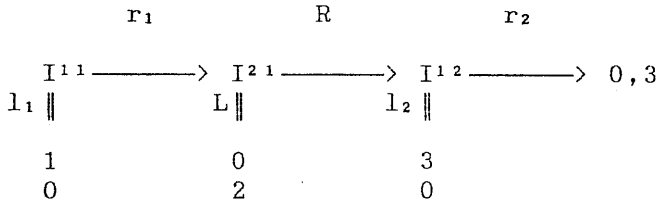
assumptions about players' beliefs need obtain. If these assumptions become common knowledge, then common knowledge of rationality follows, and we have an inconsistency as indicated by Reny. Indeed, the paradoxical conclusion is reached that *common knowledge of the theory destroys knowledge of the theory altogether*, by making it inconsistent. The problem raised by this inconsistency is, however, more general: if we want a theory of the game to include both an assumption of rationality and an assumption that this is common knowledge, do we inevitably end up with an inconsistent theory? I shall try to show that a richer theory of the game can contain both assumptions. Such a theory includes a model of belief revision specifying how the players would change their beliefs in various hypothetical situations, as when confronted with evidence inconsistent with formerly accepted beliefs [Bicchieri: 1988a].[2] A model of belief revision is especially needed in extensive form games, since a specification of the solution requires a description of what agents expect to happen at information sets that will never be reached in equilibrium play. The central idea is that a player's equilibrium strategy must prescribe a rational choice of action in all possible occurrences, including those ruled out by some putative equilibrium. But what it is rational to do at information sets off the equilibrium path depends on how a player is going to explain the fact that a deviation occurred. A model of belief revision should provide such an explanation.

In what follows I shall consider two cases: (i) that in which the players have common knowledge of the rules for belief revision but no common knowledge of their beliefs; and (ii) that in which both the rules for belief revision and players' beliefs are common knowledge. In both cases, the backward induction solution obtains.

Backward Induction

The games I am going to discuss are finite, two-person extensive form non-cooperative games of perfect information. A non-cooperative game is a game in which no precommitments or binding agreements are possible. By 'extensive form' is meant a description of the game indicating the choices available to each player in sequence, the information a player has when it is his turn to move, and the payoffs each player receives at the end of the game. Perfect information means that there are no simultaneous moves, and that at each point in the game it is known which choices have previously been made. According to the classical

theory [Kuhn: 1953], any such game has a unique solution. Take as an example the following game:



I^{ij} denotes the j-th information set ($j \geq 1$) of player i (i=1, 2). Since there is perfect information, I^{ij} is a singleton set for every i and j. Each player has two pure strategies: either to play left, thus ending the game, or to play right, and allow the other to make a choice. The game starts with player 1 moving first. The payoffs to the players are represented at the endpoints of the tree, the upper number (and the leftmost at the last branch) being the payoff of player 1, and each player is assumed to wish to maximize his expected payoff. The game is played sequentially, and at each node it is known which choices have been previously made. Player 1, at his first node, has two possible choices: to play l_1 or to play r_1. What he chooses depends on what he expects player 2 to do afterwards. If he expects player 2 to play L at the second node with a high probability, then it is rational for him to play l_1 at the first node; otherwise he plays r_1. His conjecture about player 2's choice at the second node is based on what he thinks player 2 believes would happen if she played R. Player 2, in turn, has to conjecture what player 1 would do at the third node, given that she played R. Indeed, both players have to conjecture each other's beliefs and conjectures at each possible node, until the end of the game. The classical solution of such games is obtained by backward induction as follows: at node I_{12} player 1, if rational, will play l_2, which grants him a maximum payoff of 3. Note that player 1 does not need to assume 2's rationality in order to make his choice, since what happened before the last node is irrelevant to his decision. Thus node I_{12} can be substituted by the payoff pair (3, 0). At I_{21} player 2, if rational, will only need to believe that 1 is rational in order to choose L. That is, player 2 need consider only what she expects to happen at subsequent nodes (i.e., the last node) as, again, that part of the tree coming before is now strategically irrelevant. The penultimate node can thus be substituted by the payoff pair (0, 2). At node I_{11}, rational player 1, in order to

choose l_1, will have to believe that 2 is rational and that 2 believes that 1 is rational (otherwise, he would not be sure that at I_{21} player 2 will play L). From right to left, nonoptimal actions are successively deleted (the optimal choice at each node is indicated by doubling the arrow), and the conclusion is that player 1 should play l_1 at his first node.

In the classical account of such a game, this represents the only possible pattern of play by rational players. Note, again, that specification of the solution requires a description of what both agents expect to happen at each node, were it to be reached, even though in equilibrium play no node after the first is ever reached. Thus the solution concept requires the players to engage in hypothetical reasoning regarding behavior at each possible node, even if that node would never be reached by a player playing according to the solution.

The theory of the game we have just described makes a series of assumptions about players' rationality, knowledge and beliefs, from which the backward induction (b.i.) solution necessarily follows. Let us consider them in turn. First of all, the players have to have k-th level knowledge of their respective strategies and payoffs. Second, the players must be rational, in the sense of being expected utility maximizers. Third, the players are assumed to believe each other to be rational and, depending on the length of the game, to have iterated beliefs of k-th degree about each other's rationality. It is easy to verify that in game G_1 (as in any game of perfect information) there is a belief hierarchy every two levels of which can be separated, in that there will be an action for which one level in the hierarchy will suffice, but no lower level will. At different stages of the game, one needs different levels of beliefs for backward induction to work.[3] For example, if R_1 stands for 'player 1 is rational', R_2 for 'player 2 is rational', and $B_2\ R_1$ for 'player 2 believes that player 1 is rational', R_1 alone will be sufficient to predict 1's choice at the last node, but in order to predict 2's choice at the penultimate node, one must know that rational player 2 believes that 1 is rational, i.e. $B_2\ R_1$. $B_2\ R_1$, in turn, is not sufficient to predict 1's choice at the first node, since 1 will also have to believe that 2 believes that he is rational. That is, $B_1B_2\ R_1$ needs to obtain. Moreover, while R_2 only (in combination with $B_2\ R_1$) is needed to predict L at the penultimate node, $B_1\ R_2$ must be the case at I_{11}. More generally, for an N-stage game, the first player to move will have to have a N-1-level belief that the second player believes that he is rational ... for the b.i. solution to obtain.

One property generally required of an agent's beliefs is that

they are internally consistent. Thus, for example, player i cannot believe that j is rational and not expect j to choose his best response strategy. It must be added that in game theory the notions of knowledge and belief are state-based, where the state a player is at is his information set. An agent i cannot possibly believe p at information set I^{ij} if his being at that information set contradicts p. Alternatively, one can say that p must be consistent with the information available to the player at the information set I^{ij}. For the purposes of our discussion, we require an individual's beliefs to have two properties: (a) they must be internally consistent, and (b) i's beliefs at any point in the game must be a function of his view of the history of the game up to that point.

Distributed Knowledge

It has been argued that at I_{21} it is by no means evident that player 2 will only consider what comes next in the game [Binmore: 1987; Reny: 1987]. Reaching I_{21} may not be compatible with backward induction, in the sense of not being consistent with the above stated assumptions about players' beliefs and rationality. Indeed, I_{21} can only be reached if 1 deviates from his equilibrium strategy, and this deviation stands in need of explanation. When player 1 considers what player 2 would choose at I_{21}, he has to have an opinion as to what sort of explanation 2 is likely to find for being called to decide, since 2's subsequent action will depend upon it. Obviously enough, different explanations lead to different expected payoffs from playing the same choice leading to I_{12}.

What player 2 infers from 1's move, though, *depends on what she believes about player 1*. Up to now, we know that different players need different levels of beliefs for the b. i. solution to obtain. More precisely, the theory of the game assumes the players to make use of all of the propositions in '$R_1 \wedge R_2 \wedge B_2 R_1$' (which stands for '1 is rational and 2 is rational and 2 believes that 1 is rational'). It might be asked whether it makes a difference to the backward induction solution that the theory's assumptions about players' beliefs are known to the players. This might mean several things. One the one hand, the theory's assumptions can be 'distributed' among the players, so that not all players have the same information. That is, the beliefs attributed to the players by the theory are differentially distributed among them, as opposed to the case in which all players share the same beliefs. In this latter case, all players are endowed with

the same information. In both cases, the players do not know what the other players know (i.e., which are the other players' beliefs).

We may imagine the players being two identical reasoning machines programmed to calculate their best action which are 'fed' information in the form of beliefs. The machines are capable of performing inferences based on the available information, which consists of 'beliefs' about the other machine. A machine can be fed more, less, or the same information than another machine. Let us look at the case in which the beliefs 'fed' to each machine are the minimal set consistent with successful backward induction. Each player can infer about the other what her own beliefs allow her to, and no more. In fact, *this allocation of beliefs is implicit in the classical solution*. Assuming the players to be rational, beliefs are thus distributed:

Player 1 believes : *Player 2 believes* :

R_2 R_1

B_2 R_1

Evidently, player 2 *does not know* that 1 believes R_2, nor that 1 believes that she believes R_1. But since she believes R_1, she plays L at I_{21} [we assume that if a belief is consistent with reaching an information set, then that belief is maintained]. Given her belief that player 1 is rational, the only inference that 2 can draw from being at I_{21} is that player 1 chose r_1 either because he does not believe that player 2 is rational (i.e., $\sim B_1R_2$), or does not believe that 2 believes that he is rational (i.e., $\sim B_1B_2R_1$) or any combination thereof. Thus 2's beliefs and knowledge of the game *allow* the play of r_1 by rational player 1, since her belief that 1 is rational is not contradicted by reaching information set I_{21}. It follows that 2's rational response is still L. Player 1 does not know what 2 believes, but he believes R_2 and B_2R_1; therefore he should play l_1, whereas 2 does not know that he should choose it. It must be noticed that the conclusion follows both from players' rationality and from distributed knowledge of beliefs (and iterated beliefs) among them.

Common Knowledge

Intuitively, one might expect that the more the players know about the theory of the game, the more enhanced their (and the theory's) predictive capability would be. That is, the more the

players know about each other's knowledge and beliefs, the more they become able to fully replicate the opponent's reasoning. Yet, as Reny has shown, assuming the players to have common knowledge of the theory of the game makes the theory inconsistent at some information set [Reny: 1987]. In fact, Reny's result can be obtained even if one assumes that the players only have common knowledge of the theory's hypotheses regarding their beliefs [Bicchieri: 1989]. That is, all players know that all players believe that '$R_1 \wedge R_2 \wedge B_2 R_1$' is true, and they all know that they all know,... ad infinitum. From this assumption, common knowledge of rationality naturally follows. To see why common knowledge of beliefs leads to an inconsistency, let us detail what each player knows under this condition:

Player 1 knows :	*Player 2 knows*
$B_2 \; R_2$	$B_1 \; R_1$
$B_2 \; R_1$	$B_1 \; R_2$
$B_2 \; B_1 \; R_2$	$B_1 \; B_2 \; R_1$
.	.
.	.
.	.

To get the backward induction solution, such an infinite chain of beliefs is not even necessary. The players need only both believe that '$R_1 \wedge R_2 \wedge B_2 R_1$' is true. Thus player 1 should choose l_1 at information set I_{11}. Suppose that I_{21} were reached. Player 2 believes $R_1 \wedge B_1 R_2 \wedge B_1 B_2 R_1$. But, since her node has been reached, one or more of the conjuncts must be false. If it is the case that $\sim B_1 B_2 R_1$, then rational 1 may have played r_1, but in this case player 2 will respond with L. If $\sim B_1 R_2$, it is also the case that rational 1 may have played r_1, and again 2 will respond with L. Only were $\sim R_1$ to be assumed would 2 respond to r_1 with R.

But can $\sim R_1$ be assumed? Both players are rational; each knows he is rational, but does not know that the other is rational. So much is postulated by the theory of the game. If common knowledge of beliefs is the case, each player will know that the other believes himself rational. Whereas one cannot be rational without knowing it (there is no such thing as 'unconscious' rationality), does knowing that somebody believes himself rational mean knowing that he is in fact rational? In general, the fact that somebody believes that p in no way implies that that person knows p, for one may know only true things, but believe many falsehoods. If p were false, one could not know that p, but still believe that p is the case.

Yet the implicit and explicit assumptions that game theory makes about the players allow one to infer from i's belief that he is rational that i knows that he is rational. Let us consider them in turn. (i) Throughout game theory, it is implicitly assumed that the meaning of rationality is common knowledge among the players. The players know that being rational means maximizing expected utility, and know that they know,..... Were a player to use another rule, he would know he is not rational (as one cannot be 'unconsciously' rational, one cannot be 'unconsciously' not rational). A fortiori, he could never believe he is rational. Still, it is possible that a player is rational but lacks the calculating capabilities required to compute the equilibrium solution (or solutions), or has a mistaken perception of his payoffs and strategies. In this case, knowing that i is rational is not sufficient to predict his moves. We thus need to add the following clauses: (ii) the players are perfectly able to follow through the reasoning process, as complicated as it may be, and (iii) the players have k-th level knowledge of the complete description of the game. This means each player knows his (and the other's) payoffs and strategies, and knows that the other knows,... And this rules out misperception.

If common knowledge of their respective beliefs thus implies common knowledge of rationality, it follows that $\sim R_1$ cannot be assumed. But then, of course, player 2 cannot assume 1 not to believe R_2, nor can she believe that 1 does not believe $B_2\ R_1$. If rationality is common knowledge, the conjunction $R_1 \wedge R_2 \wedge B_2 R_1$ must be true, but then a deviation from equilibrium is inconsistent with rationality common knowledge. Player 1 knows that 2 will reach some conclusion, but he is unable to tell which one. Indeed, *allowing common knowledge of beliefs destroys common knowledge of rationality*.[4]

A theory of belief revision

It has been suggested that the only solution to the above problem is to abandon either the assumption that the players are rational, or the common knowledge assumption [Reny: 1987]. As I have shown elsewhere, common knowledge of beliefs (and therefore of rationality) is neither necessary nor sufficient for the backward induction solution to obtain [Bicchieri: 1989]. The problem raised by Reny, however, is more general. Is it really the case that, in the kind of games we are considering, common knowledge of rationality *always* leads to inconsistencies? In what follows, I argue that it need not, in that a richer theory of the

game can contain both players' rationality and common knowledge of it.

We may start by considering that when a player has to choose a move, he will ask himself what the other player would do in response to his choice. In G_1, for example, player 1 must know that 2 would respond with L to r_1 in order to choose l_1. To be able to decide how another player would react to one's choice, each player has to ask how another player would explain an unexpected move or, in other words, how a deviation from the equilibrium strategy would be interpreted. Before the game is played, each player will have a model of the game which includes some beliefs about the other player's beliefs and rationality. Given that players' mutual beliefs are not common knowledge, we know that each player will be able to infer from his model the unique equilibrium solution.

Asking what would happen were a deviation from equilibrium to occur means asking - from the viewpoint of the model of the game - a counterfactual question. In order to answer it, a player has to revise his original model so as to accommodate the antecedent of the counterfactual, and then look for the consequent in the revised model. There will in general be many ways to revise one's model. The theory of belief revision proposed here fulfills two desirable requirements: (i) the original model should be revised so as to maintain consistency, and (ii) the revised model should seek to explain deviations in a way that is compatible with players' rationality. These requirements, it must be noted, capture some important features of the idea of 'rationalizability'. (i) corresponds to the requirement that a player should not entertain a belief that does not reach the information set at which he is [Pearce: 1984, p. 1041]. (ii) is analogous to requiring that if an information set can be reached without violating any player's rationality, then the conjecture held at that information set must not attribute an irrational strategy to any player [Pearce, ibid.]. Since I have extensively discussed this theory of belief revision and its implications for game theory elsewhere [1988a], I shall only outline here the bare essentials.

The best known model of belief change is Bayesian conditionalization. But conditionalization only applies to changes of beliefs where a new sentence is accepted which is not inconsistent with the initial corpus of knowledge, while the type of belief change we are discussing involves a sentence that is not a serious possibility, given the background knowledge of the players. Thus the type of belief change we are discussing requires one to accept *less* than one did before in order to

investigate some sentence that contradicts what was previously accepted. Such changes are fairly common in hypothetical reasoning, and have been variously called "question opening" [Harper: 1977] and "belief contravening" [Rescher: 1964; Levi: 1977]. Gärdenfors [1978, 1984] has proposed a model of belief change which specifically focuses on the factors governing changes of beliefs when earlier accepted beliefs are contracted in order to add a new belief which is inconsistent with the previous belief system held by the agent.

We assume each player i to start 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. L is assumed to be closed under the standard truth-functional connectives and to be ruled by a logic *L* which contains all truth-functional tautologies and is closed under Modus Ponens.

The weak rationality conditions that M_i^0 has to satisfy are spelled out in Gärdenfors [1978, 1984]. Such a set, it must be added, 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. Of course, some of the accepted sentences may be probabilistic judgments, such as probability assignments to other players' types or strategies. What matters is that in an agent's state of belief all such assignments will have probability one.

The initial model of the game M_i^0 (i=1,2) will contain statements describing the rules of the game, the players' strategies and payoffs, and statements to the effect that the above statements are common knowledge. Since we do not want beliefs to be common knowledge, let us assume that the following set of sentences is also part of the model, but that the model contains no sentence saying that the following sentences are common knowledge:

(i) the players always play what they choose at all nodes;
(ii) '$R_1 \wedge R_2 \wedge B_2 R_1$' at all nodes;
(iii) player 1 chooses l_1;
(iv) player 1 plays l_1;

To decide what to do, 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 were played. In order to consider the possibility of a deviation occurring, the player has to eliminate from M_i^0 all those beliefs which 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.

There will in general be many ways to fulfill this requirement. For example, since (iv) is implied by the conjunction of (i) and (ii), eliminating (iv) implies eliminating the conjunction of (i) and (ii). This means eliminating (i), or eliminating (ii), or eliminating both. In turn, since (ii) is itself a conjunction, eliminating it means eliminating any number of its conjuncts. Besides maintaining consistency, it seems reasonable to require belief changes to satisfy a further rationality criterion: that of avoiding unnecessary losses of information. In this case, the players face two "minimal" choices compatible with the elimination of (iv): either (i) and (iv) are eliminated, or (iv) and one of the statements in (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 a piece of information is more 'useful' than another does not mean that it is better confirmed, more probable or even more plausible. Following Levi [1977, 1979], we may distinguish between *degrees of acceptance* and *degrees of epistemic importance*. 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-making 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 Mi0 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 instead means eliminating first all those sentences which possess lower informational value [Levi: 1977, 1979; Gärdenfors: 1984]. It must be noted that introducing 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 well be the same.

The changes of beliefs we are discussing involve accepting a sentence the negation of which was earlier accepted; such belief-contravening changes can be better analyzed as a sequence of contractions and expansions, as has been suggested by Levi [1977]. Let us denote the *contraction* of a belief set M with respect to a sentence A by M_{-A}. The *expansion* of a belief set M with respect to a sentence A will be denoted by M_{+A}. The minimal set of weak rationality conditions that both contractions and expansions of belief sets have to satisfy are discussed in Gärdenfors [1984].

Suppose $\sim A \varepsilon M$. Then in order to add a belief contravening statement A, one will first contract M with respect to $\sim A$, and then expand M_{-GA} by A. By definition, $M_A=(M_{-GA})_{+A}$. We may call the revised belief set M_{-GA} a *counterfactual change* of M. Indeed, when a player asks himself "if there were a deviation from the equilibrium strategy l_1, then..." he is asking a counterfactual question (from the viewpoint of the model of the game he starts with), answering which means first contracting and then expanding his 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 the revised belief set M_A which results from minimally changing M to include A (i.e., iff $B \varepsilon M_A$).

It remains to be established how the revised belief set is to be constructed. Supposing 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 will want M_{-GA} to be as large a subset of M as possible. Gärdenfors has suggested that we define M_{-GA} as *maximally consistent* with A in relation to M iff for every $B \varepsilon M$ and $\notin M_{-GA}, (B \rightarrow \sim A) \varepsilon M_{-GA}$. Thus, if M_{-GA} were expanded by B, it would entail $\sim A$ [Gärdenfors: 1984]. Still there might be many subsets of M which are maximally consistent with A.[5] This means that the players may not revise their beliefs in the same way, thus ending up with the same solution.

Wanting the ordering of maximally consistent contracted belief sets to be complete provides a good reason to introduce further restrictions. Another reason for supplementing the cri-

terion of maximal consistency is the following: suppose that the 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 AvB and Av~B. If M is minimally contracted with respect to A, then either AvB or Av~B will belong to M_{-A}. If M_{-A} is expanded by ~A, $(M_{-A})_{+GA}$ will contain either B or ~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.[6]

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. It remains to be established how one can order sentences according to their informational value or epistemic utility. If we admit that all the sentences in an agent's belief set are equally acceptable, it will be impossible to discriminate among them in terms of probability, evidential support, or plausibility. When judging the loss of informational value caused 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. As Isaac Levi puts it, informational value is "partially dependent on the demands of the inquiries which X regards as worth pursuing at the time. Thus, the simplicity, explanatory power and the subject matter of the hypotheses contribute to their informational value" [Levi: 1984, p.169].

Informational value, in this interpretation, is a pragmatic concept. Depending on the context, certain statements will be less vulnerable to removal than others, and in any context it will generally be possible to order the statements with respect to their epistemic importance. I shall assume the order of epistemic importance to be complete and transitive.[7] A rational player will thus modify his beliefs according to the following rules [Bicchieri: 1988a]:

R1. Any revised belief set should satisfy weak rationality criteria [Gärdenfors: 1984],
R2. From the set M_{-GA} of all maximally consistent contractions of M with respect to ~A, select the subset M^*_{-GA} of the 'most epistemically important' belief sets with the aid of the criterion of minimum loss of informational value,[8]
R3. The new contracted belief set M_{-GA} should include all the sentences which are common to the elements of M^*_{-GA}, i.e., $M_{-GA}=\cap M^*_{-GA}$,[9]

R4. Expand the belief set M_{-GA} thus obtained by A.

It must be noticed that while R1 corresponds to the weak rationality criteria imposed on belief sets, R2 involves a stronger, substantive rationality criterion. It implies, for example, that it is always possible to 'objectively' define relative epistemic importance, however pragmatic and context-dependent it may be. In any given game, the ordering of sentences with respect to epistemic importance must be unique, or the players may never get to converge to the same interpretation of a deviation from equilibrium. R2 says that a criterion of epistemic importance 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 which are common to the 'most important' belief sets. *We assume these rules to be common knowledge among the players.*

If we return to our example, we can imagine player 2 deciding how to contract her original model M^2_0 with respect to sentence (iv) in order to retain consistency. If M^2_0 is retracted according to R2, she is left with several maximally consistent belief sets: M_1={(ii) and (iii)}; M_2={(i) and '$R_1 \wedge R_2$'}; M^3={(i) and '$R_1 \wedge B_2 R_1$'}; M^4={(i) and '$R_2 \wedge B_2 R_1$'}.

To complete the ordering, she has to assess whether one of the contractions entails a greater loss of informational value than the others. 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 which is optimal with respect to her revised belief set.

M_1 entails substantial informational loss, since eliminating (i) introduces an ad hoc element into the explanation of behavior. Retaining the assumptions that player 1 is rational and chooses to play the equilibrium strategy means explaining a deviation as the effect of a random mistake (indeed, systematic mistakes would be incompatible with rationality). Thus even if player 1 were to make a long series of mistakes, these would be interpreted as random and uncorrelated, and each one would have to be separately explained. Since an arbitrary pattern is made compatible with rational behavior, this explanatory strategy undermines the strength of a principle of rationality.

Contractions M_3 and M_4 involve an even greater loss of informational value, since in both cases it is assumed that one of the two players does not believe the other to be rational. If rationality is abandoned, predictability is lost, too. M_2, on the

contrary, retains both the assumption that both players are rational and the behavioral regularity (i). If M_2 is expanded with respect to ~(iv), player 2 will interpret a deviation by 1 as an intentional action, compatible with 1 being rational. Player 2 will keep believing that 1 is rational, that 1 believes that she is rational, and that 1 does not make mistakes. Player 1 deviates *because* he does not believe that 2 believes he is rational, hence 2 will respond with L.

Even if we assume the players to have common knowledge of R1–R4, they will not attain common knowledge of the revised model they will both adopt. This happens because, even if R1–R4 are common knowledge, *it is not common knowledge that 2 believes 1 to be rational*, since this is not required by R1–R4. Since $B_2\ R_1$ is not common knowledge, it can only be common knowledge that, *were 2 to believe that 1 is rational*, her revised belief set would be M_2. But, as far as 2 knows, 1 might not believe $B_2\ R_1$. Player 1, in turn, believes $R_2 \wedge B_2 R_1$, but he does not know whether 2 believes that he believes $R_2 \wedge B_2 R_1$. If 2 were to believe that 1 believes $R_2 \wedge B_2 R_1$, she would play L, and if she were not to believe that 1 believes $R_2 \wedge B_2 R_1$, she would still retain the belief that 1 is rational, and thus play L. Therefore 1's conclusion is to play l_1, which is precisely what the backward induction theory predicts.

Suppose now that *both* R1–R4 and M_i^0 are common knowledge among the players. Now of course the revised belief set will be common knowledge among them, too. Does this make the theory of the game inconsistent at some information set?

Since M_i^0 is now common knowledge, a new ordering of the contracted belief sets with respect to epistemic importance is necessary. If M_2 is adopted, it is the case that $\sim B_1 B_2\ R_1$, which means that rational 1 has played r_1, and player 2 will respond with L. If M_3, then $\sim B_1 R_2$, which also means that rational 1 has played r_1, and again 2 will respond with L. In both cases this conclusion is common knowledge, which makes 1's deviating from l_1 *incompatible with his being rational*. The same is obviously true for contraction M_4. All these contractions involve the same loss of informational value, since upholding one of them would imply that at information set I_{21} player 2 would have the following pair of inconsistent beliefs:

$B_2 R_1 \wedge B_2(R_1 \rightarrow \sim B_2 R_1)$. If the second belief is true, it is not possible that 2 believes 1 to be rational, since that very belief implies that 1 is not rational, contrary to what 2 believes. Maintaining one of the above contractions would thus render the theory inconsistent at node I_{21}

The contraction involving the least loss of informational

value is now M_1, since eliminating a behavioral regularity (i.e., 'the players always play what they choose to play at all nodes') is better than having to abandon rationality. Indeed, if one of the other contractions were adopted, either there would be an inconsistency in the theory of the game, or the assumption of rationality would have to be abandoned. The belief revision model therefore recommends choosing M_1. Since this is common knowledge, it is also common knowledge that 2 will respond to a deviation from equilibrium with L, and therefore 1 will have no incentive to deviate.

Carnegie Mellon University

* A version of this paper was presented at the Conference on Theoretical Aspects of Reasoning about Knowledge, Pacific Grove, March 1988. Financial support from NSF grant SES 87-10209 is gratefully acknowledged.

NOTES

1. While the first definition of common knowledge is found in Lewis [1969], its application to game theory is due to Aumann [1976]. For recent work on common knowledge, see Brandenburger and Dekel [1985], and Tan and Werlang [1986].
2. The importance of modeling the process of belief revision has been explicitly recognized by Pearce, when stating that "The possibility of collapsing series of choices into timeless contingent strategies must not obscure the fact that the phenomenon being modeled is some sequential game, in which conjectures may be contradicted in the course of play" [1984: p. 1041].
3. The language in which we express game theoretic reasoning is a propositional modal logic for m agents. Starting with primitive propositions p, q, ... , more complicated formulas are formed by closing the language under negation, conjunction, and the modal operators B_1... Bm and K_1... Km [Hintikka: 1962].
4. From the result that common knowledge of rationality mates the theory of the game inconsistent, Reny has inferred that the players may have an incentive to create an environment in which common knowledge is no longer possible [Reny: 1987]. However, if rationality were common knowledge it would also be common knowledge that player 2 would not

know how to interpret a deviation on the part of 1. That is, it would be common knowledge that the theory of the game is inconsistent, and therefore that 'anything can happen'. Thus a 'deviation' on the part of player 1 is not necessarily interpreted as a signal by 2 [Bicchieri: 1988b].

5. The maximally consistent contractions have been subsequently called "maxichoice contractions" by Alchourron, Gärdenfors and Makinson [1985].
6. This difficulty is pointed out in Gärdenfors [1984] and in Alchourron, Gärdenfors and Makinson [1985].
7. A similar proposal is found in Gärdenfors [1984].
8. Being able to order sentences by epistemic importance does not give an ordering of sets of sentences. Since the sets we are considering are finite, though, we can identify the informational value of a set of sentences with the informational value of the sentence which is the conjunction of all the sentences contained in the set. I am grateful to Michael Bacharach for pointing this out to me.
9. This type of contraction function is outlined in Gärdenfors [1984] and its properties are spelled out in Alchourron, Gärdenfors and Makinson [1985].

REFERENCES

C. E. Alchourron, P. Gärdenfors and D. Makinson: 1985, 'On the Logic of Theory Change: Partial Meet Contraction and Revision Functions', *The Journal of Symbolic Logic* 2, 510- 530.

R. Aumann: 1976, 'Agreeing to Disagree', *The Annals of Statistics* 4, 1236-1239.

D. Bernheim: 1984, 'Rationalizable Strategic Behavior', *Econometrica* 52, 1007- 1028.

C. Bicchieri: 1988a, 'Strategic Behavior and Counterfactuals', *Synthese* 76, 135-169.

------ 1988b, 'Common Knowledge and Backward Induction: A Solution to the Paradox', in M. Vardi (ed.), *Theoretical Aspects of Reasoning About Knowledge*, Morgan Kaufman, Los Altos.

------ 1989, 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge', *Erkenntnis*, 30, 69-85.

A. Brandenburger and E. Dekel: 1985, 'Hierarchies of Beliefs and Common Knowledge', Research Paper No. 841, Graduate School of Business, Stanford University.

P. Gärdenfors: 1978, 'Conditionals and Changes of Belief', *Acta Philosophica Fennica XXX*, 381- 404.

------ 1984, 'Epistemic Importance and Minimal Changes of Be-

lief', *Australasian Journal of Philosophy* 62, 136- 157.
W. Harper: 1977, 'Rational Conceptual Change', *PSA 1976*, vol.2, Philosophy of Science Association, 462- 494.
J. Hintikka: 1962, *Knowledge and Belief*, Cornell University Press, Cornell.
H. W. Kuhn: 1953, 'Extensive Games and the Problem of Information', in H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games*, Princeton University Press, Princeton.
I. Levi: 1977, 'Subjunctives, Dispositions and Chances', *Synthese* 34, 423- 455.
------ 1979, 'Serious Possibility', *Essays in Honor of Jaakko Hintikka*, Reidel, Dordrecht, 219- 236.
------ 1984, *Decisions and Revisions*, Cambridge University Press, New York.
D. Lewis: 1969, *Convention*, Harvard University Press, Cambridge.
D. Pearce: 1984, 'Rationalizable Strategic Behavior and the Problem of Perfection', *Econometrica* 52, 1029- 1050.
P. Reny: 1987, 'Rationality, Common Knowledge, and the Theory of Games', *Mimeo*, The University of Western Ontario.
N. Rescher: 1964, *Hypothetical Reasoning* , North Holland, Amsterdam.
T. Tan and S. Werlang: 1986, 'On Aumann's Notion of Common Knowledge - An Alternative Approach', *Working Paper* 85 - 26, University of Chicago.