CRISTINA BICCHIERI

SELF-REFUTING THEORIES OF STRATEGIC INTERACTION: A PARADOX OF COMMON KNOWLEDGE*

INTRODUCTION

Game theoretic reasoning is sometimes strikingly inconsistent with observed behavior, or even with evidence from introspection. Famous examples of such inconsistency are the finitely repeated Prisoner's Dilemma game and Selten's Chain Store Paradox (Selten, 1978). In both cases, some plausible solutions run counter to game theoretic reasoning and appear to point to the inadequacy of the game theoretic notion of rationality in capturing important features of human behavior. These considerations do not apply to artificial settings only: in a wide range of ordinary social interactions it does not pay to be (or look) too rational. As Goffman puts it, expert poker players sometimes discover that one can lose the game because of playing too well (Goffman, 1969). In international conflicts, it may well pay to be thought of as a 'mad dog'. And in the above quoted games, all or some of the players involved get a lower payoff than they would were they to play in less than a rational manner. All attempts to explain the emergence of cooperation in a finitely repeated Prisoner's Dilemma, as well as that of 'reputation effects' in the Chain Store story, have required either some version of 'imperfect rationality' (Selten, 1978) or a change in the structure of the game, such as assuming altruism (and thus changing the payoffs), or imposing incomplete information at the beginning of the game (Kreps et al., 1982).

The attempts at bridging the gap between the correct but implausible game theoretic results and the plausible (but unexplained) observed outcomes have assumed the logical inescapability of the classic game theoretic solution. However, as much as it is true that only under certain conditions can one behave as a 'mad dog', or in less than a rational manner, it is also the case that the standard game theoretic solution can only obtain under special conditions. These conditions, as I shall argue, pertain to the players' knowledge of the theory of the game (i.e., the theory of backward induction in finite games of perfect information). Here by 'theory' I mean the following

Erkenntnis **30** (1989) 69–85.

things: a set of assumptions about the players' rationality and their beliefs about each other's rationality, a specification of the structure of the game, of the players' strategies and payoffs and the hypothesis that structure, strategies and payoffs are known by the players. From these assumptions the unique equilibrium solution is derived.

How much do the players need to know about the game for them to successfully complete the reasoning required of them, and infer the unique solution? Intuitively, one might expect that the more the players know about each other, the easier it should be for them to replicate each other's reasoning and to predict each other's play. To prove otherwise is the aim of the present paper. In fact, it can be shown that in order for the backward induction solution to obtain, the players must have some knowledge of the theory's assumptions, but no common knowledge of them.[1] The paradoxical conclusion we reach is that *common knowledge of the theory of the game makes the theory inconsistent.*

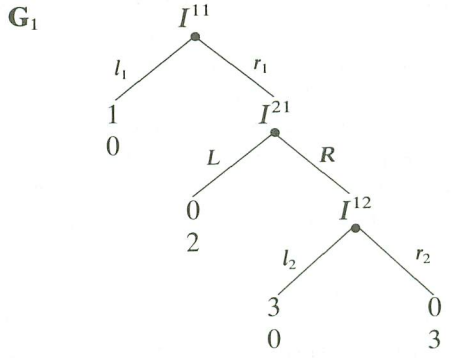
An obvious requirement a theory of the game has to satisfy is that it be free from contradictions at every information set (Reny, 1987). If a player were to find herself at an information set with which the theory of the game she is using is inconsistent, she would be deprived of a theory upon which to base her decisions. This would leave the other players (and the game theorist) without a theory, too, since they would become unable to predict what she will do, and would therefore be unable to decide what to do themselves. Consistency and predictability are strictly related. In perfect information games, it is enough to assume the players to have common knowledge of the assumptions regarding their beliefs to make the assumptions inconsistent with some information set. This conclusion implies that once common knowledge of the theory of the game is assumed, it is no longer necessary to change the structure of the game, or to reject the traditional definition of rationality, to allow for different solutions. Indeed, the existence of common knowledge makes deviations from the classical solution plausible, and compatible with individual rationality.

Knowing the disruptive effects of common knowledge of beliefs over the theory of the game, the players may have an incentive to manipulate knowledge in order to attain higher payoffs. Manipulation of knowledge, in this context, means that one or more players would communicate their beliefs to the rest of the players. At the end of the paper, I shall only briefly suggest how this result can bear upon the

cooperative solution in the finitely repeated Prisoner's Dilemma and the 'reputation effects' of the Chain Store Paradox.

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 backward induction 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 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.

BELIEFS, ITERATED BELIEFS, COMMON KNOWLEDGE

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 (if the game has $k+1$ stages). 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.

One property generally required of an agent's beliefs is that they be 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. 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 information set must be consistent with the information available to the player at that information set.

The language in which we are going to express game-theoretic reasoning is a propositional modal logic for m agents. Starting with primitive propositions $p, q, \ldots$, more complicated formulas are formed by closing the language under negation, conjunction, and the modal operators $B_1, \ldots, B_m$ and $K_1, \ldots, K_m$ (Hintikka, 1962). The very idea of iterated beliefs, however, requires a generalization of the notion of belief from an individual i to a group G (Halpern and Moses, 1986). Let us define $\mathbf{B}_G p$ ('everyone in G believes p') in the following way: $\mathbf{B}_G p$ holds iff all members of G believe p. Formally,

$$\mathbf{B}_G p \equiv \bigwedge_{i \in G} B_i p$$

$\mathbf{B}^k_G p$, $k \geq 2$ ('p is $\mathbf{B}^k$-belief in G') is defined by

$$\mathbf{B}^1_G p = \mathbf{B}_G p,$$
$$\mathbf{B}^{k+1}_G p = \mathbf{B}_G \mathbf{B}^k_G p, \text{ for } k \geq 1$$

p is said to be $\mathbf{B}^k$-belief in G if "everyone in G believes that everyone in G believes that ... that everyone in G believes that p is the case" holds, where the phrase "everyone in G believes that" appears in the

sentence k times. Equivalently,

$$\mathbf{B}_G^k p \equiv \bigwedge_{i_j \in G,\, 1 \leq j \leq k} B_{i_1} B_{i_2} \ldots B_{i_k} p$$

There are circumstances in which we may require the agents' beliefs to be common knowledge. For example, if it is publicly announced that "I believe that you believe that I am rational", everybody will know that I believe that you believe that I am rational, and everybody will know that everybody knows that . . . , and so on ad infinitum. If $K_i p$ stands for "i knows p", let us define "everybody knows p" as

$$\mathbf{K}_G p \equiv \bigwedge_{i \in G} K_i p$$

Iterated knowledge of p can be thus expressed:

$$\mathbf{K}_G^k p \equiv \bigwedge_{i_j \in G,\, 1 \leq j \leq k} K_{i_1} K_{i_2} \ldots K_{i_k} p$$

Let us now define $\mathbf{C}_G p$('p is common knowledge in G') as follows: p is said to be common knowledge in G if p is true, and is $\mathbf{K}_G^k$-knowledge for all $k \geq 1$. In other words,

$$\mathbf{C}_G p \equiv p \wedge \mathbf{K}_G p \wedge \mathbf{K}_G^2 p \wedge \ldots \wedge \mathbf{K}_G^m p \wedge \ldots .$$

In particular, $\mathbf{C}_G p$ implies all formulas of the form $K_{i_1} K_{i_2} \ldots K_{i_n} p$, where the i_j are all members of G, for any finite n, and is equivalent to the infinite conjunction of all such formulas. Clearly, the notions of group knowledge introduced above form a hierarchy, with $\mathbf{C}_G p \supset \cdots \supset \mathbf{K}_G^{k+1} p \supset \cdots \supset \mathbf{K}_G p \supset p$.[2] A similar hierarchy is formed by the notions of group beliefs, e.g., $\mathbf{B}_G^{k+1} p \supset \cdots \supset \mathbf{B}_G p$.

It is easy to verify that in game G_1 (as in any game of perfect information) every two levels of the belief hierarchy 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. 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_2R_1. B_2R_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, $B_1B_2R_1$ needs to obtain. Moreover, while R_2 only (in combination with B_2R_1) is needed to predict L at the penultimate node, B_1R_2 must be the case at I^{11}. More generally, for an N-stage game, the first player to move will have to have the $N-1$-level belief that the second player believes that he is rational . . . for the b.i. solution to obtain.

DISTRIBUTED KNOWLEDGE AND FULL 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 a theory of 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 making the same move 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 that the players make use of all of the propositions in '$R_1 \wedge R_2 \wedge B_2R_1$' (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. On the one hand, the theory's assumptions can be 'distributed' among the players, so that not all players have the same information. That is, 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 player believes.

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 upon the available information, which consists of 'beliefs' about the other machine. A machine can be fed more, less, or the same information as another machine. Let us look first 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 his own beliefs allow her to, and no more. In fact, this allocation of beliefs is implicit in the classical solution.[3] Assuming the players to be rational, beliefs are thus distributed:

Player 1 believes:	Player 2 believes:
R_2	R_1
B_2R_1	

Evidently, 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}. Given her belief, 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 player 2's knowledge of the game and beliefs allow the play of r_1 by rational player 1, since her belief that 1 is rational is not contradicted by reaching mode I^{21} (I assume that if a belief is consistent with reaching an information set, then that belief is maintained). 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 this conclusion follows both from players' rationality and from distributed knowledge of beliefs (and iterated beliefs) among them.

It is easy to verify that, were the players to have the same beliefs, the backward induction solution would still obtain. In this case, everybody has to believe that '$R_1 \wedge R_2 \wedge B_2R_1$' is true. Given that it is redundant to have a player believe that he or she is rational or believes something (they are supposed to be rational and to know what beliefs they have), this distribution of knowledge in no way modifies the conclusion that a deviation from equilibrium is consistent with the players' beliefs (and therefore with the assumptions of the theory).[4] Thus backward induction works even if all players know the same set

of propositions. In both full and distributed knowledge, however, the players have been assumed not to know what the other believes.

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. In what follows, it will be assumed that the players have common knowledge of the theory's assumptions regarding their beliefs. That is, all players know that all players believe that '$R_1 \wedge R_2 \wedge B_2R_1$' is true, and they all know that they all know, ... ad infinitum. The paradoxical conclusion is that a theory that is common knowledge among the players becomes inconsistent.[5] To see why common knowledge of beliefs leads to inconsistency, let us detail what each player knows under this condition:[6]

Player 1 knows:	Player 2 knows:
B_2R_2	B_1R_1
B_2R_1	B_1R_2
$B_2B_1R_2$	$B_1B_2R_1$
⋮	⋮

To get the backward induction solution, such an infinite chain of beliefs is not even necessary. The players need only both believe that '$R_1 \wedge R_2 \wedge B_2R_1$' is true. Thus player 1 should choose l_1 at node I^{11}. Suppose that I^{21} were reached. Player 2 knows $B_1R_1 \wedge B_1R_2 \wedge B_1B_2R_1$. But, since the node has been reached, one or more of the conjunction's elements must be false. Common knowledge of beliefs does not allow player 2 to assume that either $\sim B_1B_2R_1$ or $\sim B_1R_2$ is the case, and this is common knowledge. The deviation can only be explained assuming that $\sim R_1$; in this case, 2 would 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 a player 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 common 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, we know that 2 cannot assume 1 not to believe R_2, nor can she believe that 1 does not believe B_2R_1. If rationality is common knowledge, the conjunction $R_1 \wedge R_2 \wedge B_2R_1$ must be true, but then a deviation from equilibrium is inconsistent with rationality common knowledge. In other words, a deviation from equilibrium leads player 2 to uphold the following pair of inconsistent beliefs at node I^{21}: $B_2R_1 \wedge B_2(R_1 \rightarrow \sim B_2R_1)$. 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. Assuming common knowledge of the theory's assumptions about players' beliefs would thus render the theory inconsistent at node I^{21}.

If a deviation from equilibrium occurs, is it really the case that 1 is

not rational, or is he only trying to cheat player 2 into responding with R? Both things are possible, and there is just no way for player 2 to rationally decide in favor of one of them. Given that this reasoning process is virtual, in the sense that it takes place in the mind of each player before the game starts, player 1 will be unable to predict what 2 would choose were he to deviate from the b.i. solution, since his very deviation would make the assumption that both believe '$R_1 \wedge R_2 \wedge B_2R_1$' to be true inconsistent with reaching node I^{21}.

Indeed, *allowing common knowledge of the theory of the game makes that theory inconsistent.* Thus, were the players to tell each other what they believe, or were a public announcement made stating that they both believe $R_1 \wedge R_2 \wedge B_2R_1$, the theory of the game would immediately become inconsistent.[7]

MANIPULABILITY OF KNOWLEDGE

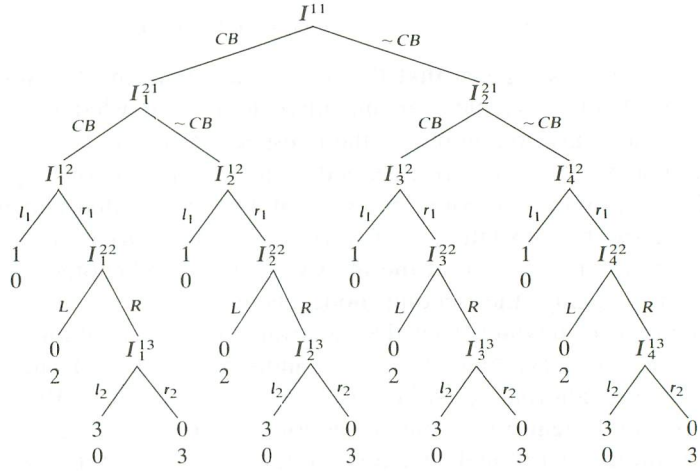
These findings suggest that the players, knowing this 'common knowledge effect', may have an incentive to use knowledge strategically (i.e., they may communicate their respective beliefs to each other). If the players' beliefs were assumed to be common knowledge in game $\mathbf{G}_1$, for example, it would be common knowledge that player 2 would not know how to interpret a deviation on the part of 1. It would be common knowledge that the theory of the game becomes inconsistent upon reaching the second node, hence player 1 might have an incentive to deviate from the backward induction solution. Since the *choice* of deviating becomes indistinguishable from irrational behavior, alternative solutions are made possible. By 'irrational' behavior I mean the choice of actions which do not maximize expected utility. Irrational behavior does not correspond with erratic behavior; if it were so, one would become unable to distinguish, in a finite number of repetitions, rational from irrational behavior, since rational and random choices might happen to coincide. This fact, however, would deprive the very idea of rationality of its predictive power, since a sequence of choices that appear to be rational might just be the result of chance. In what follows, irrational behavior is understood to mean automatically playing 'right' at every node, irrespective of one's beliefs or knowledge.[8]

Suppose now that common knowledge of beliefs is not assumed (as is the case with backward induction theory). *Would the players have an incentive to make their beliefs common knowledge*? If so, how can we

incorporate this new strategic element into the structure of the game? For common knowledge of beliefs to obtain, the players have to communicate to each other their respective beliefs. Since we want this communication to be part of the game, we may think of 'belief communication' as a type of action, and thus add to the original game two possible choices for each player: to communicate beliefs (CB) or not to communicate beliefs (~CB).

The players start the game with the distributed knowledge of beliefs assumed by backward induction theory (e.g., player 1 believes the conjunction $R_1 \wedge R_2 \wedge B_2 R_1$ to be true, and player 2 believes the conjunction $R_2 \wedge R_1$ to be true). Adding the choices of communicating or not communicating their beliefs changes the original game $\mathbf{G}_1$ into the following game:

$\mathbf{G}_2$



I_p^{ij} denotes the j_p-th node ($j \geq 1, p \geq 0$) of player $i (i = 1, 2)$. The game starts with player 1 moving first: he can choose to communicate his beliefs (CB) or not to communicate them (~CB). Player 2 moves afterwards, and she, too, has the choice of communicating or not communicating her beliefs. What is communicated is a statement to the effect that a player believes something to be the case. Let us call this statement **p**; in order for **p** to become common knowledge, however, **p** must be true (otherwise what becomes common knowledge is that a statement **p** has been made, not its truth value). Thus for **p** to be common knowledge it must be evident (and common knowledge)

to both players that *player 1 would not have an incentive to communicate* **p** *if* **p** *were not true*.

Suppose player 1 says that he believes the conjunction $R_1 \wedge R_2 \wedge B_2R_1$ to be true. Let **p** stand for the statement "I believe '$R_1 \wedge R_2 \wedge B_2R_1$' to be true". How is player 2 going to decide that **p** is true? If **p** were false, at least one of the conjuncts would be false, and the question then is whether player 1 would have any incentive to communicate **p** nonetheless. There are three possible cases:

(i) $\sim R_1$ is the case. An irrational player has no incentive to communicate that he is rational, since in any case he is going to deviate from his equilibrium strategy; indeed, even if $\sim R_1$ were false, player 1 would have an incentive to communicate that he does not believe he is rational, so as to 'cheat' player 2 into responding with R. Hence if B_1R_1 is communicated, it must be true:[9]

(ii) $\sim B_1B_2R_1$ is the case. Communicating it, together with B_1R_1, would make player 2 change her belief, were she to believe 1 to be irrational, since B_1R_1 must be true. But then a deviation from the equilibrium strategy would make the theory of the game (from 2's viewpoint) inconsistent with reaching node I_1^{21}. Hence if $B_1B_2R_1$ is told, it must be true;

(iii) $\sim B_1R_2$ is the case. Unless player 2 were to communicate that she is rational, rational player 1 would have a reason to deviate from the equilibrium strategy and no reason to tell a lie. Telling B_1R_2 and believing otherwise would not change his strategy, nor would it change player 2's strategy if 2 is irrational as 1 believes. Then if B_1R_2 is communicated it must be true.

However, the fact that one does not have an incentive to tell a lie does not make what one says necessarily true. *Communicating does not necessarily make* **p** *common knowledge*. Common knowledge of **p** is *supported* by a set of consistent beliefs that the players *might* entertain, but these beliefs are not the only possible consistent beliefs about what player 1 would communicate. That is, if player 2 were to argue along the lines depicted in points (i)–(iii), she would believe **p** to be true, and conversely, if player 1 were to believe 2 thus believes, he would believe that 2 believes that **p** is true. It is therefore possible for 1 to communicate **p** and to believe that 2 holds beliefs that make **p** common

knowledge. Even if common knowledge of **p** does not necessarily follow from communicating **p**, *knowing that there exists a consistent set of beliefs supporting common knowledge of* **p** *is enough to induce player 1 to deviate from his equilibrium strategy.*

Does player 1 have an incentive to keep silent? The answer is negative. Since at the start of the game the players are endowed with distributed knowledge of beliefs, player 1 *does not know* how player 2 is going to interpret his silence. Given his beliefs about player 2 (i.e., that 2 is rational and believes 1 to be rational) 1 can expect 2 to respond to a deviation with L. Hence it is always better for 1 to communicate **p**, but player 2 does not know that since, by assumption, 2 does not know 1's beliefs at the start of the game.

What about player 2? It is easy to verify that, whatever 1 does, 2 can either keep silent or communicate her beliefs. If **p** is communicated and 2 keeps silent, a deviation from equilibrium can occur, but a deviation can occur if 2 communicates her beliefs, too. It might be thought that, were 1 to keep silent, player 2 would have an incentive to communicate a false belief: that is, that she is not rational. In this case, though, both players would know 2 has an incentive to tell a false belief, so that 2's purpose is defeated.

The idea that knowledge can be thus manipulated has interesting applications. In the finitely repeated Prisoner's Dilemma, it is well known that cooperation can result when the players' rationality is not common knowledge among them (Kreps et al., 1982). In this case both players have a motive to deviate from their classical equilibrium strategies, since the expected payoff of, say, a tit-for-tat strategy is greater than what can be obtained by using the b.i. solution. Instead of assuming incomplete knowledge of each other's rationality on the part of the players, the solution proposed by Kreps et al. can be rationalized in terms of a larger game in which the players have the choice to communicate their beliefs. If they do, this makes it possible for both to deviate from the non-cooperative equilibrium. The same considerations apply to Selten's Chain Store Paradox (Selten, 1978), with the difference that only the Chain Store benefits from deviating from the classical solution. If so, the Chain Store should rationally choose to communicate its beliefs, so as to make the theory of the game inconsistent and to prevent the competitors from using it.

The point here is not that of *predicting* cooperation or reputation effects. They may or may not occur, depending on such elements as

the players' psychological propensities, their previous histories and experience, and their capability to interpret the other player's moves. The relevant consideration is that *alternative solutions can be shown to be fully compatible with the players' rationality*. For example, in a Prisoners' Dilemma game repeated 100 times, player 1 may decide to cooperate (C) in the first round, and for the next rounds $N = 2, \ldots, T < 100$ to choose C in period N unless player 2 chose to defect (D) in period $N-1$. For rounds $N > T$, he will always defect, regardless of the other player's choice. Were 2 to play D in period $N-1$, 1 will respond with D in period N. He may keep playing D until player 2 chooses C, and then play C again. However, he may signal to player 2 his willingness to cooperate by returning to play C immediately after he played D in the previous round. Or they may alternate in playing C and D. In general, it can be shown that a cooperative pattern is better for both, and that there are several cooperative equilibria. The precise solution, however, is impossible to predict, both because there are many possible patterns of cooperation, and because each player will probably make a different 'guess' as to the magnitude of T. Indeed, each cooperative equilibrium assumes the players to have common knowledge of their possible strategies, and of the probabilities each assigns to the other's strategies.

In the Chain Store Paradox, backward induction dictates that the chain store play cooperatively with every competitor, and that each competitor enter the local market. While in the short run the cooperative response is more advantageous, we know that in the long run the aggressive response may be a better choice, since it would discourage possible competitors from entering the market. If the game has N periods and there are N competitors, one for each period, the chain store may decide to play aggressive (A) for $N-T$ periods in response to a competitor entering the market, and to cooperate (C) in the remaining T periods. The pattern of play is, however, unpredictable. It depends on such elements as how successful the threat is, and on players' expectations as to the size of T. In this case, too, the 'aggressive' solutions depend on the players having common knowledge of the strategies and probabilities.

In both cases, alternative solutions can be rationally justified without having to introduce notions such as bounded rationality, altruism, or incomplete information. Provided the players are able to evaluate the strategic effect of introducing common knowledge of the

theory of the game (through common knowledge of beliefs), they can plausibly decide to communicate their beliefs. This move will open up different patterns of play that provide them with higher expected utilities.

NOTES

* I am grateful to Tommy Tan and Philip Reny for helping me appreciate the importance of common knowledge in games, and to Jon Elster and Michael Woodford for many useful comments. Financial support from National Science Foundation Grant SES 87–10209 is gratefully acknowledged.

[1] For the players to have *common knowledge* that p means that, not only does everyone know that p is true, but everyone knows that everyone knows, everyone knows that everyone knows that everyone knows, and so on ad infinitum.

[2] $\mathbf{K}_G p \supset p$ because $Kp \supset p$ (i.e., one cannot know something which is not true).

[3] This point is seldom recognized by game theorists. Reny discusses the importance of players' knowledge of the theory's assumptions for the b.i. solution to obtain; however, he seems to assume that without common knowledge of backward induction "a more precise treatment of how it operates is called for", since "if one of the players is not familiar with backward induction logic, then he may not play according to its prescriptions. In this case other players (even those familiar with backward induction) may rationally choose not to play according to the prescriptions of backward induction" (Reny 1987, p. 48). As I show, common knowledge of the theory of the game is neither necessary nor sufficient to obtain the b.i. solution.

[4] It is easy to verify that both sentences "all players believe $R_1 \wedge R_2 \wedge B_2 R_1$" and "player 1 believes $R_2 \wedge B_2 R_1$ and player 2 believes R_1" translate into $\mathbf{B}_G R_2 \wedge \mathbf{B}_G R_1 \wedge \mathbf{B}_G^2 R_1$ (where $G = 1, 2$).

[5] Phil Reny has independently shown that assuming the players to have common knowledge of their respective rationality makes the theory inconsistent at some information set (Reny, 1987). I obtain the same result assuming that the players have commmon knowledge of the theory's hypotheses regarding their beliefs. From this assumption, common knowledge of rationality naturally follows.

[6] If we substitute the sentence "all players believe the statement '$R_1 \wedge R_2 \wedge B_2 R_1$' to be true" with p, common knowledge that p corresponds to the infinite conjunction: $p \wedge K_1 p \wedge K_2 p \wedge K_1 K_2 p \wedge K_2 K_1 p \wedge \ldots$.

[7] I have shown elsewhere that a richer theory of the game (i.e., a theory that includes a model of belief revision) can accommodate common knowledge of rationality without giving rise to inconsistencies (Bicchieri, 1988a, 1988b).

[8] There are of course many alternative possible interpretations of irrational behavior. For simplicity, I assume the players to have common knowledge that they can be either rational or irrational in the special sense specified above.

[9] In the previous section, it has been shown that if a player believes he is rational, he also knows he is rational, thus if player i says $B_i R_i$ it means i knows he is rational and therefore he is rational.

REFERENCES

Aumann, R. J.: 1976, 'Agreeing to Disagree', *The Annals of Statistics* **4**, 1236–9.
Bicchieri, C.: 1988a, 'Strategic Behavior and Counterfactuals', *Synthese* **75**, 1–35.
Bicchieri, C.: 1988b, 'Common Knowledge and Backward Induction: A Solution to the Paradox', in M. Vardi (ed.), *Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufman Publishers, Los Altos.
Binmore, K.: 1987, 'Modeling Rational Players', Part I, *Economics and Philosophy* **3**, 179–214.
Goffman, E.: 1969, *Strategic Interaction*, University of Pennsylvania Press, Philadelphia.
Halpern, J. and Y. Moses: 1986, 'Knowledge and Common Knowledge in a Distributed Environment', *Research Report*, IBM Almaden Research Center.
Hintikka, J.: 1962, *Knowledge and Belief*, Cornell University Press, Cornell.
Kreps, D., P. Milgrom, J. Roberts and R. Wilson: 1982, 'Rational Cooperation in the Repeated Prisoner's Dilemma', *Journal of Economic Theory* **27**, 245–52.
Kuhn, H. W.: 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.
Lewis, D.: 1969, *Conventions*, Harvard University Press, Cambridge.
Luce, R. and H. Raiffa: 1957, *Games and Decisions*, Wiley, New York.
Reny, P.: 1987, 'Rationality, Common Knowledge, and the Theory of Games', *mimeo*, Department of Economics, University of Western Ontario.
Rosenthal, R.: 1981, 'Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox', *Journal of Ec. Th.* **25**, 92–100.
Schelling, T.: 1960, *The Strategy of Conflict*, Oxford University Press, New York.
Selten, R.: 1978, 'The Chain-Store Paradox', *Theory and Decision* **9**, 127–59.
Tan, T. and S. Werlang: 1986, 'On Aumann's Notion of Common Knowledge – An Alternative Approach', *Working Paper* 82–26, University of Chicago.

Manuscript submitted 25 January 1988
Final version received 31 March 1988

The University of Chicago
Center for Ethics, Rationality and Society
5828 South University Avenue
Chicago, IL 60637
U.S.A.

Department of Philosophy
University of Notre Dame
Notre Dame, IN 46556
U.S.A.