CRISTINA BICCHIERI

STRATEGIC BEHAVIOR AND COUNTERFACTUALS*

ABSTRACT. The difficulty of defining rational behavior in game situations is that the players' strategies will depend on their expectations about other players' strategies. These expectations are beliefs the players come to the game with. Game theorists assume these beliefs to be rational in the very special sense of being *objectively correct* but no explanation is offered of the mechanism generating this property of the belief system. In many interesting cases, however, such a rationality requirement is not enough to guarantee that an equilibrium will be attained. In particular, I analyze the case of multiple equilibria, since in this case there exists a whole set of rational beliefs, so that no player can ever be certain that the others believe he has certain beliefs. In this case it becomes necessary to explicitly model the process of belief formation. This model attributes to the players a theory of counterfactuals which they use in restricting the set of possible equilibria. If it were possible to attribute to the players the same theory of counterfactuals, then the players' beliefs would eventually converge.

1. MUTUAL RATIONAL BELIEFS

In interactive contexts, such as those treated in game theory, what it is rational to do depends on what one expects that other agents will do. Could we take these expectations as given, the problem of strategy choice would be simple: an agent would choose the strategy which maximized his payoff under the assumption that all other agents act in accordance with his expectation. If the agents' reciprocal expectations are not given, then the problem arises of what expectation is to be entertained by a rational player who expects the other players to act rationally, and to entertain similar expectations about him and the other players. The problem has been appropriately termed by Harsanyi one of 'mutual rational beliefs' (Harsanyi 1965).

There are games in which this problem is absent, and what it is rational to do is straightforward; for example, there may exist a 'dominant strategy', which yields a better outcome than any other strategy, whatever the other players do. But in general this is not the case.

When an explicit treatment of expectations is required, game theorists assume that players are equipped with subjective probability distributions over the other players' choices and are *practically rational*

Synthese **76** (1988) 135–169.

in that they act so as to maximize expected utility relative to these. These subjective probability distributions are beliefs the players come to the game with, but game theory does not explain them as the result of some rational procedure such as, for example, Bayesian updating from primitive priors.

Rather, it is assumed that the players' beliefs are *in equilibrium* at the start of the game, which customarily means that all players' beliefs about the behavior of the rest are consistent and are common knowledge among the players.[1] Hence an equilibrium belief has the following stability property: were a player to take the other players' beliefs as given, he would not want to change his belief. It must be noted that equilibrium beliefs are not synonymous with consistent beliefs: a set of beliefs might be consistent, but it might be the case that, were the players endowed with common knowledge of each other's beliefs, they would have an incentive to revise them. While an equilibrium belief is always a consistent belief, the reverse need not be true, as the following example makes clear.

Suppose two players face the following normal form game G_1:[2]

G_1

		II: **l**	II: **r**
I	**L**	1, 1	1, 1
	R	−1, −1	2, 0

Both players are rational, in that they maximize their expected utilities, and believe each other to be rational. Yet rationality only dictates the choice of a strategy which is a best response to a conjecture about the other's choice, not to the other's actual choice. Thus were player I to believe, for whatever reason, that player II will play **l** with probability 1, then his best response will be to play **L**. Similarly, if II were to believe that I will play **L** with probability 1, then her best response is to play **l**. Since I believes II to be rational, expecting II to play **l** with

probability 1 means assuming that strategy **l** is a best response, on the part of player II, to some conjecture about player I's action, which in turn must be a best response to his conjecture about player II's action, and so on.

The combination of strategies that these beliefs support is a Nash equilibrium, which is defined as a set of strategies, one for each player, such that for each player his strategy is a best response to the strategies of the other players. The game G_1 has two pure strategy equilibria: (**L**, **l**) and (**R**, **r**), and each is supported by a range of beliefs sufficient to attain it. In order to attain an equilibrium, however, not any combination of beliefs will do. For example, if I were to believe that II will play **l** with probability 1/2, and II were to believe that I will play **L** with some probability less than one, I will play **L** and II will play **r**, which is not an equilibrium. It is easy to see that in this case the players' beliefs can be consistent, but are not in equilibrium: if I believes II to play **l** with probability 1/2, then I must also believe that II believes I to play **L** with probability 1, and even if II's *actual* belief is inconsistent with I's second-order conjecture about II's belief, each player's conjectures may be perfectly consistent. More generally, an *internally consistent conjecture* can be defined as follows: if player *i* has a certain belief about player *j*, then he must believe his belief is among the possible beliefs which *j* has about *i* himself (Tan and Werlang 1986b).

Consistency among beliefs does not mean that they are in equilibrium.[3] Take as an example the following beliefs: player I believes that II plays **r** with probability 2/3, while player II believes that I plays **R** with probability 1. It is easy to verify that these beliefs are consistent, and indeed they support the equilibrium (**R**, **r**). Yet, if it were common knowledge that player II expects I to play **R** with probability 1, then I would not retain his belief that II will play **r** with probability 2/3. I would now believe that II will play **r** with probability 1. Were the players' initial beliefs to become common knowledge, then not only they would undergo revision, but the players would also know how they have been revised, and know that they know... ad infinitum.

In game G_1 there are two possible configurations of equilibrium beliefs: either both players assign probability 1 to (**L**, **l**) being played, or they assign probability 1 to (**R**, **r**). Yet Nash equilibria, far from being a consequence of rationality, may arise from implausible expectations: even if **l** is a weakly dominated strategy for player II, (**L**, **l**)

remains an equilibrium, since the definition of equilibrium beliefs supporting it does not rule out 'implausible' beliefs.[4]

It is important to realize that what have just been termed 'equilibrium beliefs' are all that is meant by 'rational beliefs' in game theory: rationality is conceived as a *relation* between beliefs, not a *property* resulting from the procedure by which the beliefs are obtained. Only if the beliefs of the players are both consistent and common knowledge among them are they rational in the above sense. This explains why game theorists are quite agnostic as to the rationality of individual, isolated beliefs; such beliefs are, a priori, neither rational nor irrational, only acquiring one or the other property when confronted with other individuals' beliefs. Rationality is an attribute of belief ensembles rather than individual beliefs. However, the equilibrating mechanism generating this property of belief ensembles is left unexplained or, in other words, we do not know how beliefs come to be common knowledge.

This admittedly limited definition of mutually rational beliefs would be completely satisfactory were game theory just bound to define what an equilibrium is and the conditions which make it possible. Given the definition of a Nash equilibrium, it follows that each equilibrium is supported by a set of mutual rational beliefs, in the sense defined above. Yet normative game theory's aim is to prescribe actions that will bring about an equilibrium, which means providing a *unique* rational recommendation on how to play. Indeed, if the task of the theorist were limited to pointing to a set of rational actions, the players might never succeed in coordinating their actions, since different agents might follow different recommendations. Thus a unique rational action for every player must be recommended, together with a unique belief about the behavior of other players justifying it.

Furthermore, the unique rational belief in question must be shown to be necessarily held by any player who is rational in an appropriate sense. For to say that a Nash equilibrium is supported by a configuration of equilibrium beliefs is still just part of a description of that equilibrium, and gives remarkably little guidance as to what to expect the players to do. In fact, I shall argue that so minimal a requirement of belief-rationality is often insufficient to guarantee that an equilibrium will occur. In game G_1, for example, the fact that the players have common knowledge of their respective beliefs that, say, I will play **L** with probability 1 and II will play **l** with probability 1, does not guarantee that II will not randomize between **l** and **r**. Common

knowledge of beliefs, in other words, does not allow the players to *deduce* each other's strategy.

Even in the presence of a unique configuration of rational beliefs, a player might not be sure what to expect the other players to do, in the sense that there might not be any reason to expect the other players to choose their equilibrium strategies. It is enough to look at a simple case such as a two-person noncooperative game with a unique Nash equilibrium (without dominant strategies), to realize that even here game theory is not at all explicit in describing how the players come to choose their equilibrium strategies (Spohn 1982; Bacharach forthcoming).[5] Let us look, as an example, to a modified version of G_1:

G_2

		II: **l**	II: **r**
I	**L**	1, 1	1, 1
	R	2, −1	0, 0

The unique Nash equilibrium is (**L**, **r**); it is supported by a set of equilibrium beliefs (I believes that II will play **r** with probability 1, and II believes that I will play **L** with probability 1), but common knowledge of their respective beliefs and of their both being rational players does not allow player I to deduce that player II will not randomize between **l** and **r**.

Each player must not just know that a combination of strategies is an equilibrium (even a unique equilibrium), he must also know that every other player is aiming at it. But each is aiming at that equilibrium only if each has a reason to expect the others to be aiming at it... in an infinite regress of expectations. It is not enough that both players assign probability 1 to their respective Nash strategies and that their assessments be coherent and common knowledge if there is no story of how they rationally came to hold such beliefs. The missing part in the transition from assuming rational behavior and beliefs to deriving an equilibrium configuration of actions is precisely an account of what grounds players' reciprocal expectations, a description of the process that leads players' beliefs to converge. This I shall call *the problem of justifying equilibrium play*.

This is a very general problem, and I do not intend to attack it here. I shall instead treat all those cases in which an equilibrium can be deduced, and thus justified, from an assumption of rational beliefs and behavior. One such assumption is that the players, besides being practically rational, possess common knowledge of each other's practical rationality. In very simple cases, common knowledge of practical rationality will be sufficient to identify the equilibrium the players will aim at as well as to justify equilibrium play.[6]

Yet there are many more games in which common knowledge of the fact that each player is an expected utility maximizer is consistent with the existence of more than one set of rational expectations and behaviors, so that knowing how to behave rationally will depend on knowing the other players' expectations. Such is the case when there exist multiple equilibria, each of which is supported by a set of equilibrium beliefs, but none of which is clearly 'better' than the others. Unless the gratuitous assumption is made that the players have a priori common knowledge of their respective beliefs, we only know that – *were* their beliefs common knowledge – a Nash equilibrium would obtain. Some mechanism to attain common knowledge of beliefs must thus be specified.

In the absence of direct knowledge of each other's beliefs, the players will have to infer them. Inferring other players' beliefs is not a simple matter, though. This is not only because, in games with simultaneous moves, all actions are *possible actions*, and the beliefs supporting them *possible beliefs* (unless there are strongly dominated strategies), so that a player's reasoning about another player's action and beliefs is hypothetical. It can often happen that an opponent's possible strategy choice can be justified by many different beliefs on his part, in which case a player should attempt at demarcating between plausible and implausible beliefs. A player, that is, should be able to decide whether another player's possible belief is plausible, or rational in a more substantive sense.

Belief-rationality, in this context, cannot just be a relation between a belief and other beliefs. What is needed is an interpretation of rationality as a property of individual beliefs which are the outcome of a rational process, and a description of such a process of belief formation. If there were a unique rational process of belief formation, any rational player could be expected to adopt it. Then assuming common knowledge of rationality would allow each player to infer any

other player's belief, and know that the other players can similarly infer the beliefs he entertains. This I shall call *the problem of attaining common knowledge of mutual beliefs*.

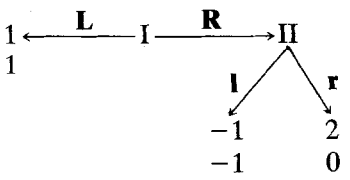
Were the players to attain common knowledge of their mutual expectations, *a fortiori* they would be able to identify their equilibrium strategies and to justify equilibrium play. The problem of justifying equilibrium play and that of attaining common knowledge of mutual beliefs are indeed, in the cases I am going to discuss, one and the same problem.

Explicitly considering the process of belief formation, however, has a price, since it introduces an element which is external to the structure of the game. For example, there are cases in which knowledge of previous actions on the part of another player alone allows one to infer the unique rational belief supporting a given choice of the other player. Since this knowledge is such as to permit a unique rational inference to be drawn, common knowledge of the players' rationality and of this piece of information will suffice to ground common knowledge of mutual beliefs.

In all cases in which it is not possible to draw a unique inference, further conditions need to be imposed on the process of belief formation. For example, there might exist several alternative processes of belief formation, giving rise to many different plausible beliefs. In this case, the players should possess some means to assess the merits of the different processes of belief formation. If it were possible to show that one of these processes is the uniquely rational one, rational players should be expected to adopt it. Then common knowledge of rationality would suffice to bring about common knowledge of beliefs. In the latter case, a solution to the problem of attaining common knowledge of mutual beliefs, and thus of justifying equilibrium play, depends upon the existence of a unique rational process of belief formation.

While the problem of attaining common knowledge of beliefs is common to both normal form and extensive form games,[7] it can be argued that extensive form games are not equivalent to their strategic form (Selten 1965, 1975; Kreps and Wilson 1982).[8] For our purposes, it is sufficient to note that an analysis based on the normal form representation ignores the role of beliefs, especially those beliefs about possible actions off the equilibrium path. This means that in the normal form the restrictions on belief required by rationality are not sufficient to rule out implausible beliefs. Since the extensive form

shows the sequential nature of the game, this representation allows one to model the process of belief formation, and thus to impose more restrictions on beliefs than those based on rationality alone. In order to see the difference between normal form and extensive form, let us consider the extensive form representation of game G_1:[9]



We know that one Nash equilibrium is (**L**, **l**). Since player I moves first, he will play **L** only if he expects II to choose **l**. Player II, in turn, chooses **l** *only if she does not have to choose*: were I to play **L**, there would indeed be no further choice to make for player II. Yet, why should I expect II to play **l**? If they were to agree to play the equilibrium (**L**, **l**) and a deviation were to occur, II would never respond with **l**. This means that, if we investigate the (**L**, **l**) equilibrium in terms of the players' decision trees, there are no probability assessments that one can put in II's decision tree that will lead her to play **l**. Knowing this fact, player I should never play **L**, since by playing **R** he will always get a higher payoff. It follows that even if in the normal form the equilibrium (**L**, **l**) is supported by 'rational' beliefs, it is not 'rationalizable' in terms of the players' decision trees, since it involves unreasonable expectations.

A satisfactory theory of belief formation must then tell us how the players would change their beliefs in various hypothetical situations, such as when confronted with evidence inconsistent with formerly accepted beliefs.[10] In what follows, I shall outline a theory of out of equilibrium belief revision restricted to extensive form games, in that only this representation allows us to model belief revision in deviant situations. The theory of belief revision presented here is based on a principle of minimum loss of informational value (Levi 1979, 1984; Gärdenfors 1984). The players revise their beliefs eliminating first those beliefs which have lower informational value, where I assume a criterion of informational value that induces a complete and transitive ordering of the sentences contained in a belief set. A principle of minimum loss of informational value, among other things, provides an

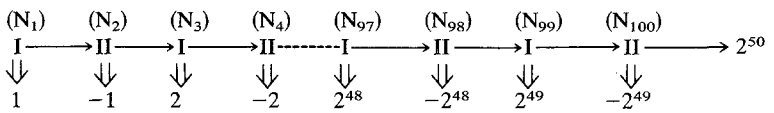
epistemic ground for maintaining – as Pearce and Bernheim do – that if an information set can be reached without violating the rationality of any player, then the agent's conjecture must not attribute an irrational strategy to any player (Bernheim 1984; Pearce 1984). Provided that common knowledge both of the initial belief set and of the rules for belief revision is assumed, a criterion of equilibrium selection which eliminates intuitively implausible equilibria is specified.

The general thesis I shall defend is that in many interesting cases the assumption that the only imputs needed for obtaining a solution are 1) knowledge of the game being played, and 2) self-evident principles of rationality which do not make use of any knowledge, will not be enough to guarantee that an equilibrium will be attained.

2. REASONING FROM PRACTICAL RATIONALITY

A crucial assumption of game theory is that players believe each other to be rational. This belief, in turn, is not very helpful unless it is accompanied by each player's belief that all the players believe each other to be rational, and so on. It is seldom acknowledged by game theorists that without such an iterative mutual rationality assumption there would hardly be anything rational to do.

The following example is meant to show that even a classical and relatively unproblematic equilibrium solution, and the method used to attain it, completely depend for their validity on assuming a finite number of levels of belief in each other's practical rationality on the part of the players. I shall argue that in a two-person zero-sum game, where each player has a unique maximin strategy, and this pair of strategies is the unique Nash equilibrium for the game, other patterns of play could be equally justified if the players do not believe each other to be rational.[11]



Players I and II have two strategies: to play "across", or to play "down" and thus end the game. They play sequentially, and at each node it is known which choices have been previously made. The numbers represent player I's payoffs; player II's payoffs are exactly the

negative of these (it is a "zero-sum game"), and each player is assumed to wish to maximize his expected payoff.

Player I, at his first node, has two possible choices: to play "down" or to play "across". What he chooses depends on what he expects player II to do afterwards. If he believes that there is a high probability that player II will play "down" at the second node, then it is rational for him to play "down" at the first node; otherwise he plays "across". His conjecture about player II's choice at the second node is based on what he thinks player II believes would happen if she played "across". Player II, in turn, has to conjecture what player I would do at the third node, given that she played "across". Indeed, both players have to conjecture each other's beliefs and conjectures at each possible node, until the end of the game.

The classical equilibrium solution of such games is obtained by backward induction. If the last node were to be reached, that part of the tree not coming after it has become strategically irrelevant; therefore the decision at that node should be based solely on what comes after it. Player II can play "down", and get 2^{49}, or play "across", and lose 2^{50}. Rationality dictates playing "down" at the last node. At the penultimate node, player I need consider only what he expects to happen at subsequent nodes (i.e., the last node) as, again, the part of the tree coming before is now strategically irrelevant. If he expects II to play at the last node as just described, his rational choice at the penultimate node is to play "down" (and receive 2^{49}) rather than "across" (and expect to receive -2^{49}). 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 I should play "down" at his first node.

It must be noted that at different stages of the game, one needs different levels of beliefs for backward induction to work. For example, if R_I stands for 'player I is rational', R_{II} for 'player II is rational', and B_IR_{II} for 'player I believes that II is rational', R_{II} alone will be sufficient to predict II's choice at the last node, but in order to predict I's choice at the penultimate node, one must know that rational player I believes that II is rational, i.e., B_IR_{II}. B_IR_{II}, in turn, is not sufficient to predict II's choice at node N_{98}, since II will also have to believe that I believes that she is rational. That is, $B_{II}B_IR_{II}$ needs to obtain. Moreover, while R_I only (in combination with B_IR_{II}) is needed to predict "down" at the penultimate node, $B_{II}R_I$ must be the case at

N_{98}. 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 backward induction solution to obtain.

According to the classical account of such a game, this represents the only possible pattern of play in equilibrium. Note, again, that specification of the equilibrium 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 equilibrium 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 the equilibrium strategy.

If, for example, player I were to find himself at node N_5, the classical view tells us that he would dismiss player II's preceding action as irrelevant to his present decision. Dismissing II's preceding action as irrelevant is equivalent to interpreting II's deviation from equilibrium (playing "across" at N_4) either as the sign that II does not believe that I believes that . . . that II is rational, or as evidence that II does not believe that I is rational, or as a mistake on the part of II. In the classical solution, all these hypotheses are compatible with I's believing that II is rational, and do not force a revision of that belief. Therefore player I will play "down". Player II, in turn, is endowed by the theory with several layers of beliefs at node N_4. For example, she has a 96th-level belief that I believes that she is rational, in addition to believing that I is rational, too. Even if II does not know what I's beliefs are at N_5, she can infer, from what she believes about player I, that he will play "down" at the next node. This means that she should play "down" at N_4. Repeating this reasoning for each node, one concludes that the rational solution is for I to play "down" at the first node.[12]

Suppose now that the players do not believe in their mutual rationality. Then it is by no means obvious that, at whatever node a player imagines finding himself, he has to conclude that the rational solution is to play "down". Is there a good reason why player I, at node N_5, should interpret past deviations as suggested above? Are those interpretations about player II's deviation still justified? After all, II might be making systematic mistakes; she might just be an automaton that always plays "across". Or she might be a sophisticated player, feigning incompetence only to induce player I to play "across" a few more times, so as to raise her payoff. Taking these hypotheses

seriously might induce player I to deviate from his equilibrium strategy, and play "across" at node N_5. His deviation would be perfectly justified; indeed, it would be the only rational choice in these circumstances. Note that if player II were to expect player I to interpret her deviation in the sense just described, it would be rational for her to deviate from her classical strategy, too – she can bluff at least some of the time, playing "across" in order to benefit from I's interpretation of her behavior as that of an automaton.

In fact, one can show that another kind of equilibrium is possible for this game, if player I assigns a positive probability to the hypothesis that II is an automaton and this probability is common knowledge. The equilibrium is of the following sort. For the first several nodes of the game (how many depends on how large a probability I assigns ex ante to II's being an automaton), both players play "across" at each node. As the end of the game approaches, II begins to randomize her behavior, sometimes playing "down" and sometimes "across".[13] This is because, as the end nears, I will cease to be willing to play "across" unless he has come to assign a relatively high probability to II being an automaton (if this point in the game is reached); this follows from observing II playing "across" repeatedly only if he believes that rational II would not choose to play "across" each time. In this equilibrium, there is a positive probability of all nodes of the game being reached, even when both players are rational; II may bluff all the way to node N_{100}, at least some of the time, at which point she surprises I by playing "down".

In the absence of beliefs in their mutual rationality, the players thus face an embarrassment of riches. It is possible that both are rational; it is possible that the other player is an automaton; and it is also possible that the other player is rational and exploits the former possibility. A player might play "across" by chance, by necessity, or on purpose. Depending upon the initial configuration of beliefs attributed to the players, there can be several equilibria, and since a player has no way to tell what the other player believes, there is no justification for equilibrium play.

Assuming the players to believe in their mutual rationality, in this particular case, provides a grounding for their expectations. Indeed, it is a necessary and sufficient condition for players' actions to be rational. It is necessary because in its absence it would be impossible to identify any rational choices as such, and sufficient in that it allows

each player to identify a unique rational choice. In this relatively simple case, it is just enough to assume the players to hold a N-1-level belief in each other's rationality for an equilibrium to be attained.

3. COORDINATION OF BELIEFS

In the former example, if a finite number of levels of belief in each other's rationality were assumed, knowledge of mutual expectations would immediately follow, pointing to a unique choice for each player as the rational choice. Yet there are many more cases where it is impossible to establish that there is a single rational choice. This happens even if we impute to the players an infinite number of levels of belief in each other's rationality, or even common knowledge of rationality.

The problem is that knowing other players to be rational, and knowing that they know, and so on, might not be enough to point to a unique strategy for each player, if there is no univocal interpretation of what is the rational thing to believe, and thus to do. Let us consider as an example a simple coordination problem, where there are two possible 'meeting points', or equilibria, and the problem of the agents is that of arriving at one of them. Take this two-by-two normal form game:

		II	
		L	**R**
I	**l**	2, −2	−1, −3
	r	−1, −2	0, −1

The only pure strategy equilibria of this game are (**I**, **L**) and (**r**, **R**). In deciding which one is more likely to be aimed at, we have to reconstruct the reasoning of the players and their system of beliefs. We assume that they know each other's possible strategies and payoffs and rationality, and that this is common knowledge. However, this mutual knowledge is not enough to guarantee that they will reach an equilibrium. Or, in other words, nothing guarantees that their mutual beliefs will be in equilibrium. If player II thinks it is likely that I

chooses strategy **l**, then it makes sense for her to choose **L**. This, however, only makes sense if I has reason to think that II thinks it is very likely that he chooses **l**; then he may expect her to choose **L**, and so choose **l**. But II will think that I will probably choose **l** only if she can reasonably believe him to have a reason to do so.

This type of game has been of interest to philosophers, too, in connection with the emergence of norms or conventions (Lewis 1969; Gauthier 1975; Ullman-Margalit 1977; Gilbert 1981, 1983a). With the notable exception of Margaret Gilbert, all of these authors have suggested that a solution to the problem of multiplicity is provided by the "salience" of one equilibrium.[14] In other words, one particular equilibrium has some characteristic distinguishing it as special; however, salience can provide I with a reason to choose it only if he has some reason to expect II to be influenced by it. Salience, in other words, provides me with a reason to choose a particular strategy only if I think you believe that salience provides me with such a reason. It gives me a reason only if it gives you a reason, but it gives each of us a reason only if it gives a reason to both (Gilbert 1983b; Bicchieri forthcoming a).

For the regress to come to an end, there must be some ground to justify our reciprocal expectations. It may of course happen that we entertain mutually consistent expectations, and therefore are able to coordinate, but this does not mean that our expectations were rational in the first place. We may have had no ground to expect what we did. Maybe we need to assume there is a rule that says that, whenever we can use this or that salience criterion, we have to use it. Or maybe it is just a convention that we use it. As I have shown elsewhere (Bicchieri forthcoming a), this solution begs the question.

As opposed to the traditional game-theoretic view that a solution has to be context-free, a solution to the above game is found by introducing some feature, external to the game, which allows the players to isolate a configuration of beliefs as more plausible than the others. The players, that is, must be endowed with some form of knowledge enabling them to draw an inference about the beliefs which ground each possible action. Considering the context in which the game is played might offer such knowledge.

Suppose both players have common knowledge of the fact that player I had the following choice: either to play game 1, or to play game 2, in which case game 1 would never be played:[15]

		II L	II R
I	**l**	2, −2	−1, −3
	r	−1, −2	0, −1

game 1

I	1, −2

game 2

When player II finds herself called to play game 1, she learns something about player I's prediction about what II would do if called to play game 1, and therefore what I plans to do once he has to choose between **l** and **r**. I's choice of game 1, in turn, can only be supported by the belief that II will replicate I's reasoning, since no other belief on the part of I is compatible with choosing game 1. Both players are aware of that, therefore I can play game 1 predicting II's belief, and thus II's choice of **L**.

Such considerations direct attention to the equilibrium (**l**, **L**), whereas if player II had been the one to choose first between the two games, by the same reasoning the only plausible equilibrium would have been (**r**, **R**). In this case, it is possible to infer players' beliefs from their actions and this process, coupled with common knowledge of each other's rationality, provides the ground for mutual knowledge of beliefs, which guarantees that coordination will be attained (Kreps and Wilson 1982).

4. MULTIPLE BELIEFS

In the previous game, additional knowledge was attained by adding a new game (in extensive form games, by adding a new strategy), and letting player II decide why rational player I did not choose to play the alternative game (or the alternative strategy). It is important to notice that we made player II ask why I chose game 1 instead of game 2, interpreting this choice as a *signal* sent by rational I to an equally rational player II.

An objection to this solution is whether is it really so evident that there always exists a unique rational inference to draw from a player's choice. The same choice, in other words, could support many different beliefs, all of them apparently compatible with a player being rational.

A typical such case is that of noncooperative games of imperfect information with multiple Nash equilibria, where each equilibrium is supported by a set of consistent beliefs.[16]

To be able to recommend a unique strategy for every player, game theorists must recommend a *uniquely rational* configuration of beliefs. To do so, not only must belief-rationality be interpreted as a property resulting from the procedure by which the beliefs are obtained: it must also be shown that there is a unique rational procedure for obtaining them.

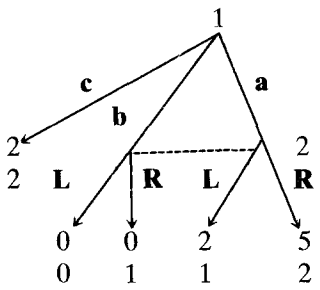
Game theorists have proposed various refinements of the Nash equilibrium concept to deal with this problem, none of them succeeding in picking a unique equilibrium across the whole spectrum of games (van Damme 1983). Within the class of refinements of Nash equilibrium, two different approaches can be 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' (Selten 1975), and Myerson's notion of 'proper equilibrium' (Myerson 1978). The alternative solution is based instead 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 (1982).

All of these solutions aim at imposing restrictions on players' beliefs, so as to obtain a unique rational recommendation as to what to believe about other players' behavior. This 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 regards the nature of the restrictions imposed on players' beliefs. In the 'small mistakes' approach, the specification of the equilibrium requires a description of what the agents expect to happen at each node (if the game is represented in extensive form), were it to be reached, even though in equilibrium play most of these nodes are never reached. The players are thus required to engage in counterfactual arguments regarding behavior at each possible node. 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 equilibrium has occurred. If, in the face of a deviation, he would still play his equilibrium strategy, then that equilibrium is 'robust', or plausible.

The following game can serve to illustrate the reasoning process through which the players come to eliminate 'implausible' equilibria:



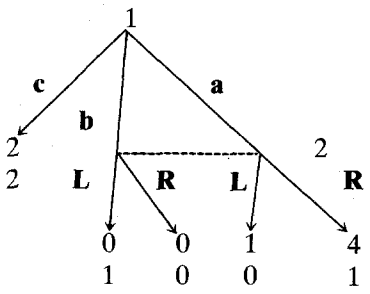
game 3

Player 1 moves first, and has to choose between strategies **a**, **b**, and **c**. If **c** is played, the game ends. However, if 1 chooses **a** or **b**, player 2 is unable to tell which of the two has been played (this *imperfect* information on the part of 2 is represented by the dotted line connecting player 2's decision nodes). The game has two Nash equilibria in pure strategies, (**c**, **L**) and (**a**, **R**). Selten rejects equilibrium (**c**, **L**) as being not sensible. 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 the players agree to play (**c**, **L**). Whether 1's choice of **c** is rational or not depends upon what he expects that 2 would do if he played **a** or **b**. 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, 2 should minimize the probability of playing **L**. This reasoning can, in fact, take place even before the unexpected node is reached, since a rational player is supposed to be able to decide what it is rational to do at every possible node, including those which would occur with probability 0 if a certain equilibrium is played.

The players are in fact required to engage in hypothetical reasoning, asking themselves what they would do if a certain node were

reached, and to understand that every information set can be reached, with at least a small probability, since it is always possible that a deviation from the equilibrium occurs by mistake. A sensible equilibrium will therefore prescribe rational (i.e., maximizing) behavior at every information set.[17] However, not all perfect equilibria are sensible, as the following example illustrates:



game 4

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 1 will make mistake **b** with a higher probability than mistake **a**, but both probabilities will be 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 limit of this proposal is that restrictions are imposed only on equilibrium beliefs, while *out of equilibrium beliefs are unrestricted*: a player will ask whether it is reasonable to believe the other player to play a given Nash equilibrium strategy, but not whether the beliefs supporting the other player's choice are rational.

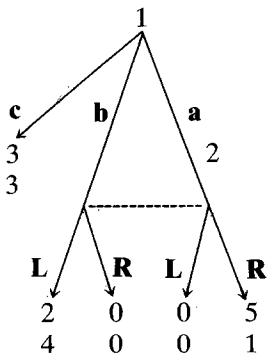
Let us compare for a moment games 3 and 4. In game 3, the equilibrium (**c**, **L**) is ruled out because player 1 cannot possibly find any out of equilibrium belief supporting it. Player 2, facing a deviation, would never play strategy **L**. In game 4, when player 1 asks whether 2 will keep playing **L** in the face of deviations, he can find some belief, on the part of 2, which would justify her playing **L**. What player 1 does not ask is whether the possible beliefs 2 might entertain about the greater or lesser likelihood of some deviation are at all justified. This, however, is a crucial question, since only by dis-

criminating between deviations (and thus between beliefs about deviations) according to their greater or lesser plausibility is it possible to restrict the set of equilibria.

To provide a satisfactory means to discriminate among equilibria, restrictions need to be imposed on all sorts of beliefs, even out of equilibrium ones. A player, that is, must be able to rationally justify to himself every belief, and expect the other players to expect him to adopt 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 all plausible deviations, defined as deviations that do not involve costly mistakes (Myerson 1978). In game 4, if player 2 were to adopt this criterion, she would assign deviation **b** a smaller probability than deviation **a**; hence she would play **R** with as high a probability as possible. This reasoning rules out equilibrium (**c**, **L**).

An easy objection to this further refinement is that a player who takes care to prevent more costly mistakes would be expected to use the same care to prevent *all* mistakes. A more serious objection is the following: while this refinement correctly attempts to restrict out of equilibrium beliefs, it suceeds in doing so only partially. There are cases in which a mistake is more costly than another insofar as the player who could make the mistake believes the other player will respond in a certain way. As the following example shows, this second-order belief stands in need of justification:



game 5

Here (**a**, **R**) and (**c**, **L**) are both 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 2 thus to play, mistake **b** would indeed be less costly than mistake **a**, in which 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 be **a** and then player 2 would choose **R**. Hence equilibrium (**c**, **L**) is highly implausible.

What these examples suggest is that for an equilibrium to be sensible *all* out of equilibrium beliefs need to be rationally justified. A player who asks himself what he would do were a deviation to occur must also find good reasons for that deviation to occur, which means justifying it by means of plausible beliefs entertained by 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 (Kreps and Wilson 1982), 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 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 maximises his expected payoff. It is easy to verify that in game 5 both Nash equilibria are also sequential equilibria. Hence such minimal rationality conditions are too weak to rule out intuitively implausible beliefs.

A possible solution 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 stresses the role of anticipated actions off the equilibrium path in sustaining the equilibrium. It thus asks the players to engage in counterfactual arguments which involve changing their original set of beliefs.[18] For the process of belief change not to be arbitrary, rationality conditions must be imposed on it. Belief-rationality, in this case, is a property of beliefs which 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 rationality would suffice to eliminate all equilibria which are only robust with respect to implausible deviations.

5. MODELING BELIEF CHANGES

In the examples discussed, elimination of implausible equilibria was attained by means of a heuristic method involving 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 isolate, whenever possible, a unique equilibrium.

My counterexamples aimed at showing 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 having been attained. Indeed, as games 4 and 5 illustrate, the players' expectations may be consistent, but they are hardly rational. This happens because the players can rationalize only *some* beliefs, in the absence of a general criterion of belief-rationality that would significantly restrict the set of plausible beliefs. This criterion, it must be added, would have the double function of allowing the players to identify a unique equilibrium and of justifying equilibrium play.

In what follows, I shall explicitly model the elimination process as a process of rational belief change on the part of the players. 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 thus defined: 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 accepting as new evidence a sentence E' whose prior probability is zero, conditionalization is not applicable as a description of belief change in this context.

Some have argued that conditionalization can be defined even if $p(E) = 0$, if one takes the conditional probability $p(H/E)$ as primitive.[19] Nonetheless, it must be noted that 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 E 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 retracted in order to add a new belief inconsistent with the previous system, and in what follows I will use his model to represent the process of belief revision taking place in the mind of the players before the start of the game.

Let us 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 $\mathcal{L}$ which contains all truth-functional tautologies and is closed under Modus Ponens. We say that a subset M of L is 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 $B \in M$.[20]

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, 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.

In attributing to each player an initial system of beliefs, I shall follow the heuristic method common to all refinements of Nash equilibrium. Each player will start by considering *only* Nash equilibria, and imagine having agreed to play one particular equilibrium, for example the equilibrium (**c**, **L**) of game 4.[21] The initial model of the game M_i^0 ($i = 1, 2$) will thus contain the rules of the game, the players' strategies and payoffs, and the following set of sentences:

(i) the players only consider Nash equilibria;
(ii) the players are rational;
(iii) the players always play what they choose;
(iv) player 1 chooses to play **c**;
(v) player 1 plays **c**;

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

To rule out implausible equilibria, 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 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 (v), 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 (v) is implied by the conjunction of (iii) and (iv), eliminating (v) implies eliminating the conjunction of (iii) and (iv). This means eliminating (iii), or eliminating (iv), or eliminating both. 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 (v): either (v) and (iv) are eliminated, or (v) and (iii).

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.[22] 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 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 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 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 the *contraction* of a belief set M with respect to a sentence A by M_{-A}. M_{-A} will satisfy the following conditions (Gärdenfors 1984):

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, $M_{-A} = M_{-B}$.

The *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 $(M_{-A})_{+A} \subseteq M$.

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). 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. By definition, $M_A = (M_{-\sim A})_{+A}$. We may call the revised belief set M_A a *counterfactual change* of M. Indeed, when a player asks himself "if there were a deviation from the equilibrium strategy $\mathbf{c}$, 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 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 the revised belief set M_A which results from minimally changing M to include A (i.e., iff $B \in M_A$).[23]

It remains to be established how the revised belief set is to be constructed. If we want the contraction of the belief set M with respect to $\sim A$ to be minimal, then in order to lose as little information as possible, we will want $M_{-\sim A}$ to be as large a subset of M as possible. Gärdenfors has suggested that we define $M_{-\sim A}$ as *maximally consistent* with A in relation to M iff for every $B \in M$ and $\notin M_{-\sim A}$, $(B \Rightarrow \sim A) \in M_{-\sim A}$. Thus, if $M_{-\sim A}$ 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.[24] Since the above inter-

pretation of 'minimal change of beliefs' is generally 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.

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. Hence, wanting the ordering to be complete provides a good reason to introduce further restrictions. Another reason for supplementing the criterion 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 $A \vee B$ 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.[25]

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.[26] A rational player will thus modify his beliefs according to the following rules:

R1 any revised belief set will satisfy C1–C13,

R2 from the set $\mathbf{M}_{-\sim A}$ of all maximally consistent contractions of M with respect to $\sim A$, select the subset $\mathbf{M}^*_{-\sim A}$ of the 'most epistemically important' belief sets with the aid of the criterion of minimum loss of informational value,[27]

R3 the new contracted belief set $M_{-\sim A}$ will include all the sentences which are common to the elements of $\mathbf{M}^*_{-\sim A}$, i.e., $M_{-\sim A} = \cap \mathbf{M}^*_{-\sim A}$,[28]

R4 expand the belief set $M_{-\sim A}$ 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. *These rules are common knowledge among the players.*

If we return to the example of game 4, we can imagine player 2 deciding how to contract her original model M_2^0 with respect to sentence (v) in order to retain consistency. If (v) is retracted according to R2, she is left with two maximally consistent belief sets: M' = (i), (ii), (iv), (vi) and M'' = (i), (ii), (iii) and (vi). In order to complete the ordering, she has to assess whether one of the two contractions entails a greater loss of informational value than the other 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 (v). Player 2 will then choose that strategy which is a best response to the revised belief set.

M' entails substantial informational loss, since eliminating (iii) 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). 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 seriously undermines the strength of a principle of rationality.

A further consideration weakens the plausibility of M'. In game 4, for example, choosing this contraction (and then expanding) implies maintaining that both equilibria (**a**, **R**) and (**c**, **L**) are sensible. However, (**c**, **L**) is sensible only if player 2 expects deviation **b** to occur more frequently than deviation **a**, and player 1 expects 2 to have such an expectation. Indeed, if 2 were not to expect 1 to expect her to assign a higher probability to deviation **b**, she could not retain sentence (iv), since in that case player 1 would not choose to play his equilibrium strategy **c**. Nothing in the original belief set M_i^0 justifies this belief on the part of player 2. Hence M' *involves arbitrary beliefs*, in that it does not require out of equilibrium beliefs to be plausible.

M'', on the contrary, by eliminating (iv), allows player 2 to interpret deviations as intentional actions, in the sense of actions intended to influence her beliefs. The revised belief set obtained from M'' implies that, facing a deviation, player 2 will interpret it as a *signal* on the part of player 1: reaching her information set, 2 deduces that 1 believes that 2 will play **R**. Given that both the initial belief set and the rules for belief revision are common knowledge, player 1 can expect 2 to know that 1 must have chosen **a**, since this is the only justifiable strategy for 1 consistent with reaching 2's information set. Player 2, in turn, knows that 1 knows that 2 knows that . . . 2 will thus interpret the deviation and respond with **R**. Since 1 can deduce this as well, he will never choose **c**. Thus (**a**, **R**) is the only sensible outcome for this game.

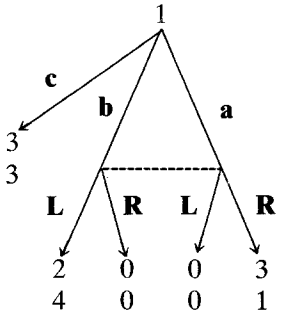
It must be noted that the revised belief set obtained from contraction M'' allows the players to deduce a unique plausible equilibrium (**a**, **R**) in the last three games, since in all of them player 1 has a reason for deviating from his equilibrium strategy **c**. This means, from the game theorist's viewpoint, that there exists a unique rational recommendation of beliefs regarding the other player's behavior. Common knowledge of mutual expectations, it must be added, justifies equili-

brium play, since rational agents will expect each other to act on the basis of their mutual rational expectations.

Is it possible, just by assuming the players to have common knowledge of M_i^0 and of R1–R4, to infer that they will attain common knowledge of the contraction strategy they will both adopt? The answer is positive only if the ordering of the sentences belong to M_i^0 with respect to epistemic importance is uncontroversial. An assumption of rationality, for example, will be difficult to give up, since it provides a very general and powerful explanatory principle. In many games, an assumption of common knowledge will be equally well entrenched, since without it even an assumption of rationality will be useless to coordinate the actions of the players.

Of the two contractions examined, M'' involves the least loss of informational value, in that it not only avoids ad hoc explanations, but extends the use of a rationality principle to out of equilibrium behavior, avoiding the implausible beliefs involved in accepting M'. In this case the application of a criterion of minimum informational loss seems straightforward, since a statement about a regularity of behavior (i.e., 'the players always play what they choose') conveys more valuable information than a statement of less general scope (i.e., 'player 1 chooses to play **c**').

We may well ask how the theory of belief revision proposed here applies to all those cases in which there is no good reason for a deviation to occur, i.e., when reaching a player's information set is not compatible with an intentional choice on the part of the other player. The following game exemplifies this case:



game 6

Suppose player 2 starts by considering equilibrium (**c**, **L**), and asks

what would happen were a deviation to occur. How will she contract her original belief set? The 'minimal' revisions will still involve deleting either sentences (v) and (iv), or (v) and (iii). Now, however, eliminating (iv) and (v), and then expanding by ~(v), implies that player 1 chose not to play **c**. Is there any other justifiable strategy on the part of 1 consistent with reaching 2's information set? Since the best outcome 1 can get by deviating is a payoff of 3, which is exactly what he could get with certainty by playing strategy **c**, an intentional deviation seems unjustified, and is indeed at odds with 1's rationality. Hence eliminating (v) and (iv), and then expanding by ~(v), involves a greater loss of informational value than the alternative revision strategy, since it makes an intentional deviation from **c** incompatible with assuming a player's rationality. The only possible contraction involves deleting (v) and (iii), while the revised belief set obtained by adding ~(v) corresponds precisely to the 'small mistakes' hypothesis.

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 play **L** instead if the chance of **b** occurring is greater than 1/5. Since the theory of belief revision adopted by the players implies that a deviation would only occur by mistake, all the conjectures entertained by player 2 are equally plausible. Hence in this case there is no way of discriminating between the three equilibria (**c**, **L**), (**c**, **R**) and (**a**, **R**), which are in fact both perfect and proper. The 'small mistakes' hypothesis is therefore a special case of our belief revision model, which only becomes plausible when deviations from equilibrium play cannot be interpreted as signals.

Finally, some analogies between this theory of belief revision and the idea of 'rationalizability' must be emphasized. The idea that a player should not entertain a conjecture that does not reach the information set at which he is (Pearce 1984, p. 1041) is similar to our requirement that, whenever considering the possibility of a move incompatible with the belief set he actually holds, a player should contract it so as to maintain consistency. A further restriction says that if an information set can be reached without violating the rationality of any player, then the conjecture held at that information set must not attribute an irrational strategy to any player (Pearce, ibid.). This principle is analogous to our emphasizing the epistemic importance of an assumption of rationality, which results in revised belief sets that

seek to explain in a reasonable way why a deviation occurred. Our results are thus similar to those obtained by applying the concept of rationalizability, with the important difference that in the case of the present belief revision model, only Nash profiles are admitted.

6. CONCLUSION

The traditional game-theoretic assumption that the only information available about other players is their current beliefs and knowledge of the game is not sufficient, in many interesting cases, to isolate a unique equilibrium. Since each equilibrium is supported by a set of rational beliefs, one has to extend the concept of belief-rationality to include a theory of how the players would revise their beliefs in deviant situations. If the players have common knowledge of an initial model of the game and of the rules for belief revision, they can attain common knowledge of their mutual expectations. Even though there are cases in which the proposed theory of belief revision will not eliminate all but one equilibria, it nonetheless offers a formalization of features of equilibria commonly perceived as salient, as well as providing a more complete characterization of rationality of beliefs.

NOTES

* I wish to thank Michael Bacharach, In-Koo Cho, William Harper, Aanund Hylland, Isaac Levi, Wolfgang Spohn, Tommy Tan and two anonymous referees for many useful comments and suggestions. 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. This concept is first introduced in D. Lewis (1969, p. 76), who shows that common knowledge is necessarily part of any convention. A formulation often used in game theory is due to Aumann (1976). T. Tan and S. Werlang (1986a) have recently offered a formalization of Lewis' concept, which they show to be equivalent to Aumann's notion.

[2] This game is taken from Kreps and Wilson (1982, p. 869).

[3] Some game theorists have explored the implications of dropping the assumption that beliefs are common knowledge. Consistent beliefs which are not common knowledge among the players have been called 'rationalizable' beliefs (Bernheim 1984; Pearce 1984). It is easy to verify that, in game G_1, all four combinations of pure strategies are rationalizable.

[4] The case of strongly dominated strategies is different; the choice of such a strategy would be plainly irrational, since a strongly dominated strategy is not a best response to any possible subjective assessment.

[5] There have been attempts on the part of philosophers to ground the Nash equilibrium concept on decision theory. W. Harper has proposed an explication of the a priori argument for Nash equilibrium by combining the concept of 'ratifiable choice' with best response reasoning (Harper 1986, 1987); B. Skyrms has explicitly modeled the equilibrating mechanism that leads Bayesian deliberators to converge to an equilibrium (Skyrms 1986, 1987a, 1987b).

[6] There are even simpler cases in which an equilibrium can be attained without common knowledge. For example, if one player has a dominant strategy, it is only needed that the other player knows that he is rational, without assuming any further level of knowledge.

[7] For a standard explanation of the terms 'normal form' and 'extensive form', see Luce and Raiffa (1957, chap. 3).

[8] The extensive form has some crucial features that the normal form lacks: some normal form equilibria require strategies that include choices and expectations that would be irrational at the points in one's decision tree where they would be made (Selten 1965, 1975). Since the game tree specifies the causal structure of the sequence of decisions and the information available at each decision point, many implausible equilibria can be eliminated through the extensive form.

[9] This game is discussed in Kreps and Wilson (1982, p. 869).

[10] 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).

[11] This game is a simplified version of Rosenthal's centipede example (1981). A similar game is also discussed in Binmore (1987).

[12] I have extensively discussed the stability of the backward induction solution with respect to deviations, and its reliance upon an assumption of distributed knowledge of beliefs on the part of the players. If common knowledge of beliefs were assumed, common knowledge of rationality would follow, and in this case the backward induction theory would become internally inconsistent (Reny 1987; Bicchieri forthcoming b).

[13] Indeed, for player II's bluff to be credible, she must seldom bluff; hence she will randomize.

[14] Such a solution was first proposed by T. Schelling (1960).

[15] This is equivalent to adding an initial strategy in the extensive form representation of the game in which player I moves first.

[16] In a game of imperfect information, the following condition obtains: at some decision point, it is not known which choices have previously been made. This is, for example, necessarily true of games with simultaneous moves.

[17] 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 (Selten 1975).

[18] Selten has explicitly discussed the role of counterfactual reasoning in decision theory and game theory in R. Selten and U. Leopold (1982).

[19] This argument originates with de Finetti. For an extensive defense of this idea, see Isaac Levi (1978).

[20] Since Gärdenfors's semantics does not refer to truth values, a 'truth functional tautology' must be intended as a purely syntactical property of formulas, implying nothing about truth values.

[21] It does not matter *which* equilibrium a player starts by considering, since he will have to repeat the reasoning for each equilibrium.

[22] I am grateful to Isaac Levi for suggesting this distinction to me.

[23] This is a formulation of the Ramsey test that does not require one to include 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. (Levi 1977, 1980). For an extensive discussion of these topics, see Gärdenfors (1978, 1986).

[24] The maximally consistent contractions have been subsequently called *maxichoice contractions* by Alchourron, Gärdenfors and Makinson (1985).

[25] This difficulty is pointed out in Gärdenfors (1984) and in Alchourron, Gärdenfors and Makinson (1985).

[26] A similar proposal is found in Gärdenfors (1984).

[27] 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.

[28] This type of contraction function is outlined in Gärdenfors (1984) and its properties are spelled out in Alchourron, Gärdenfors and Makinson (1985).

REFERENCES

Alchourron, C. E., P. Gärdenfors, and D. Makinson: 1985, 'On the Logic of Theory Change: Partial Meet Contraction and Revision Functions', *The Journal of Symbolic Logic* **2**, 510–30.

Aumann, R. J.: 1976, 'Agreeing to Disagree', *The Annals of Statistics* **4**, 1236–39.

Bacharach, M.: forthcoming, 'A Theory of Rational Decision in Games', *Erkenntnis*.

Bernheim, D.: 1984, 'Rationalizable Strategic Behavior', *Econometrica* **52**, 1007–28.

Bicchieri, C.: forthcoming a, 'Methodological Rules as Conventions', *Philosophy of the Social Sciences*.

Bicchieri, C.: forthcoming b, 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge', *Erkenntnis*.

Binmore, K.: 1987, 'Modeling Rational Players', I and II, *Economics and Philosophy* **3**, vol. 2.

van Damme, E. E. C.: 1983, *Refinements of the Nash Equilibrium Concept*, Springer-Verlag, Berlin.

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

Gärdenfors, P.: 1984, 'Epistemic Importance and Minimal Changes of Belief', *Australasian Journal of Philosophy* **62**, 136–57.
Gärdenfors, P.: 1986, 'Belief Revisions and the Ramsey Test for Conditionals', *The Philosophical Review* **XCV**, n. 1, 81–93.
Gärdenfors, P.: 1986b, 'The Dynamics of Belief: Contractions and Revisions of Probability Functions', *Topoi* **5**, 29–37.
Gauthier, D. P.: 1975, 'Coordination', *Dialogue* **14**, 195–221.
Gilbert, M.: 1981, 'Game Theory and Convention', *Synthese* **46**, 41–93.
Gilbert, M.: 1983a, 'Agreements, Conventions, and Language', *Synthese* **54**, 375–404.
Gilbert, M.: 1983b, 'Some Limitation of Rationality', paper presented at the *APA* meeting, December.
Harper, W.: 1977, 'Rational Conceptual Change', *PSA 1976*, vol. 2, Philosophy of Science Association, 462–94.
Harper, W.: 1986, 'Mixed Strategies and Ratifiability in Causal Decision Theory', *Erkenntnis* **24**, 25–36.
Harper, W.: 1987, 'Causal Decision Theory and Game Theory', *Mimeo*, University of Western Ontario.
Harper, W., R. Stalnaker and G. Pearce: 1980, *Ifs*, D. Reidel, Dordrecht.
Harsanyi, J.: 1965, 'Bargaining and Conflict Situations in the Light of a New Approach to Game Theory', *The American Economic Review* **55**, 447–57.
Kalai, E. and D. Samet: 1984, 'Persistent Equilibria', *International Journal of Game Theory* **13**, 129–44.
Kohlberg, E. and J. Mertens: 1986, 'On the Strategic Stability of Equilibria', *Econometrica* **54**, 1003–37.
Kreps, D. and R. Wilson: 1982, 'Sequential Equilibria', *Econometrica* **50**, 863–94.
Levi, I.: 1977, 'Subjunctives, Dispositions and Chances', *Synthese* **34**, 423–55.
Levi, I.: 1978, 'Coherence, Regularity and Conditional Probability', *Theory and Decision* **9**, 1–15.
Levi, I.: 1979, 'Serious Possibility', *Essays in Honour of Jaakko Hintikka*, D. Reidel, Dordrecht, 219–36.
Levi, I.: 1984, *Decisions and Revisions*, Cambridge University Press, New York.
Lewis, D.: 1969, *Convention*, Harvard University Press, Cambridge.
Lewis, D.: 1976, *Counterfactuals*, Blackwell, Oxford.
Lewis, D.: 1976, 'Probabilities of Conditionals and Conditional Probabilities', *Philosophical Review* **85**, 297–315.
Luce, R. and H. Raiffa: 1957, *Games and Decisions*, Wiley, New York.
McLennan, A.: 1985, 'Justifiable Beliefs in Sequential Equilibrium', *Econometrica* **50**, 863–94.
Myerson, R. B.: 1978, 'Refinements of the Nash Equilibrium Concept', *International Journal of Game Theory* **7**, 73–80.
Nash, J.: 1951, 'Non-cooperative Games', *Annals of Mathematics* **54**, 286–95.
Pearce, D.: 1984, 'Rationalizable Strategic Behavior and the Problem of Perfection', *Econometrica* **52**, 1029–50.
Reny, P.: 1987, 'Rationality, Common Knowledge, and the Theory of Games', *Mimeo*, The University of Western Ontario.
Rescher, N.: 1964, *Hypothetical Reasoning*, North-Holland, Amsterdam.

Rosenthal, R.: 1981, 'Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox', *Journal of Economic Thought* **25**, 92–100.
Selten, R.: 1965, 'Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit', *Zeitschrift fur die gesamte Staatswissenschaft* **121**, 301–24.
Selten, R.: 1975, 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory* **4**, 22–55.
Selten, R. and U. Leopold: 1982, 'Subjunctive Conditionals in Decision and Game Theory', in Stegmüller, Balzer and Spohn (eds.), *Philosophy of Economics*, Springer Verlag, Berlin.
Schelling, T.: 1960, *The Strategy of Conflict*, Oxford University Press, New York.
Skyrms, B.: 1986, 'Deliberational Equilibria', *Topoi* **5**, 59–67.
Skyrms, B.: 1987a, 'Deliberational Dynamics and the Foundations of Bayesian Game Theory', in J. E. Tomberlin (ed.), *Epistemology* (Philosophical Perspectives v. 2) Ridgeview, Northridge.
Skyrms, B.: 1987b, 'The Value of Knowledge', in W. Savage (ed.), *Minnesota Studies in the Philosophy of Science*, University of Minnesota Press, Minneapolis.
Spohn, W.: 1982, 'How to Make Sense of Game Theory', in Stegmüller, Balzer and Spohn (eds.), *Philosophy of Economics*.
Stalnaker, R. C.: 1968, 'A Theory of Conditionals', in N. Rescher (ed.), *Studies in Logical Theory*, Blackwell, Oxford.
Stalnaker, R. C. and R. H. Thomason: 1970, 'A Semantic Analysis of Conditional Logic', *Theoria* **36**, 23–42.
Tan, T. and S. Werlang: 1985, 'The Bayesian Foundations of Rationalizable Strategic Behavior and Nash Equilibrium Behavior', *Mimeo*, Princeton University, New Jersey.
Tan, T. and S. Werlang: 1986a, 'On Aumann's Notion of Common Knowledge – An Alternative Approach', *Working Paper* 85–26, University of Chicago.
Tan, T. and S. Werlang: 1986b, 'The Bayesian Foundations of Solution Concepts of Games', *Working Paper* 86-34, University of Chicago.
Ullman-Margalit, E.: 1977, *The Emergence of Norms*, Clarendon Press, Oxford.

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