RATIONALITY AND WELL-BEING[†]

The Traveler's Dilemma: Paradoxes of Rationality in Game Theory

By KAUSHIK BASU*

This paper presents a parable which highlights the conflict between intuition and game-theoretic reasoning. One of the basic ingredients of analysis in game theory is "backward induction"; but backward induction is also the source of some deep paradoxes (see e.g., Ken Binmore, 1987; Philip Pettit and Robert Sugden, 1989). Well-known games such as the finitely repeated "prisoner's dilemma," Reinhart Selten's (1978) "chain store," Robert Rosenthal's (1981) "centipede," and Phil Reny's (1993) "take-it-or-leave-it" highlight this conflict between backward-induction reasoning and other kinds of reasoning.

Much effort has gone into trying to solve the problem. Virtually all of these efforts exploit the extensive-form structure of the above games or the fact that they are played over time. Thus Binmore and Adam Brandenberger (1990) observe that these paradoxes arise because players in the above games can "throw surprises" on one another by deviating from the path suggested by backward induction. If all moves by all players were occuring at a point of time, "throwing surprises" would be inconsequential because it would influence no one's behavior. Reny (1993) also locates the paradox in the sequential character of these games, arguing that the problem arises because "during the course" of some plays, Bayesian rationality cannot be common knowledge.

This paper demonstrates that the above problem is deeper, because it can arise in a single-shot game. This is done by constructing a paradoxical game—the "traveler's dilemma." The backward induction in the traveler's dilemma occurs at an introspective level. The standard suggestions for battling the backward-induction paradox in, for instance, the repeated prisoner's dilemma (e.g., Cristina Bicchieri, 1989) do not seem to be possible here. Hence this problem cannot be solved by attributing unusual knowledge structures at unreached nodes.

All intuition seems to militate against all formal reasoning in the traveler's dilemma. Hence the traveler's dilemma seems to be one of the purest embodiments of the paradox of rationality in game theory because it eschews all unnecessary features, like play over time or the nonstrictness of the equilibrium.

I. The Parable

Two travelers returning home from a remote island, where they bought identical antiques (or, rather, what the local tribal chief, while choking on suppressed laughter, described as "antiques"), discover that the airline has managed to smash these, as airlines generally do. The airline manager who is described by his juniors as a "corporate whiz," by which they mean a "man of low cunning," assures the passengers of adequate compensation. But since he does not know the cost of the antique, he offers the following scheme.

Each of the two travelers has to write down on a piece of paper the cost of the

[†] *Discussant:* Thomas Schelling, University of Maryland; Hamish Stewart, University of Toronto; Robin Pope, University of New South Wales.

*Delhi School of Economics, Delhi 110 007, India. I am grateful to Cristina Bicchieri, Peter Bohm, Eric Van Damme, Martin Dufwenberg, Yvan Lengwiler, Ajit Mishra, Bob Rosenthal, Larry Samuelson, and Jorgen Weibull for useful conversations.

antique. This can be any value between 2 units of money and 100 units. Denote the number chosen by traveller i by n_i. If both write the same number, that is, $n_1 = n_2$, then it is reasonable to assume that they are telling the truth (so argues the manager) and so each of these travelers will be paid n_1 (or n_2) units of money.

If traveler i writes a larger number than the other (i.e., $n_i > n_j$), then it is reasonable to assume (so it seems to the manager) that j is being honest and i is lying. In that case the manager will treat the lower number, that is, n_j, as the real cost and will pay traveler i the sum of $n_j - 2$ and pay j the sum of $n_j + 2$. Traveler i is paid 2 units less as penalty for lying and j is paid 2 units more as reward for being so honest in relation to the other traveler.

Given that each traveler or player wants to maximize his payoff (or compensation) what outcome should one expect to see in the above game? In other words, which pair of strategies, (n_1, n_2), will be chosen by the players?[1]

In order to answer this question it is useful first to express this game as a payoff matrix. Observe that the above game could be thought of as having at least two versions, depending on whether the players can choose any real number or can choose only an integer. For most of the time I shall assume the latter, since that is where the main problem arises. When I assume the former, I shall refer to the game as the "continuum version of the traveler's dilemma."

II. The Paradox

At first sight, both players feel pleased that they can get 100 units each. To get this, each player simply has to write 100. But each player soon realizes that if the other player adheres to this plan then he can get 101 units of money by writing 99. But, of course, both players will do this, which means, that each player will in fact get 99 units. But if both were planning to write 99, then each player will reason that he can do better by writing 98; and so on. The logic is inexorable, and there is no stopping until they get to the strategy pair (2,2), that is, each player writes 2. Hence, they will end up getting two units of money each. This illustrates how backward-induction, at the level of introspection, works even in a one-shot game.

It is easy to check that all standard solution concepts predict outcome (2,2). This is the unique strict equilibrium of the game, the only Nash equilibrium, and in fact, the only rationalizable equilibrium. Yet it seems very unlikely that any two individuals, no matter how rational they are and how certain they are about each other's rationality, each other's knowledge of each other's rationality, and so on, will play (2,2). It is likely that each will play a large number in the belief that so will the other, and thereby they will both get large payoffs. At one level the traveler's dilemma shares some similarities with the Bertrand duopoly, especially one in which firms choose prices from a grid; for instance, the set of integers starting from an integer above the marginal cost and up to the monopoly price. The best-response structure of such a duopoly is similar to the best-response structure of the traveler's dilemma. However, that is where the analogy ends. In the Bertrand duopoly, if one firm chooses a price even slightly above the other's price, it earns zero profit. The penalty is nowhere nearly as severe for choosing a higher number in the traveler's dilemma. This is exactly what makes it plausible that players will choose large numbers in the traveler's dilemma. It may be possible, however, to construct a model of differentiated-products duopoly which is exactly analogous to the traveler's dilemma.

In the finitely repeated prisoner's dilemma, it has been shown that cooperation in the early games is possible if one uses the (single-shot) rationalizability criterion. In this game (2,2) is the unique rationalizable outcome. Observe also that, unlike in this

[1] This game is a generalization of the prisoner's dilemma, since, if the travelers had to confine their choice to 2 or 3, we would have exactly the prisoner's dilemma. (A different generalization of the prisoner's dilemma occurs in Basu [1994]).

game, in the centipede or the take-it-or-leave-it game the "unwanted" equilibrium is not strict.[2] Hence, in terms of formal analysis there seems to be no escape from (2,2).

But even knowing all this, there is something very rational about rejecting (2,2) and expecting your opponent to do the same. This is the essence of the traveler's *dilemma*. This is also the reason why escape routes which are made possible by allowing for irrationality or the expectation of irrationality (see e.g., David Kreps et al., 1982) are not of relevance here even though they may be important empirically.[3] It is not an empirical point that is being made here. The aim is to explain why, despite rationality being common knowledge, players would reject (2,2), as intuitively seems to be the case.

III. The Possibilities

While I am unable to resolve the paradox, what follows are some possible lines of attack. Possibility 1 suggests a rigorous resolution of the problem for a special case, to wit, the continuum version; possibilities 2 and 3 should be treated as speculative rather than formal.

Possibility 1.—The continuum version of the traveler's dilemma has an interesting way out by using an adaptation of the concept of curb sets, developed in Basu and Jorgen Weibull (1991)—curb being an acronym for "closed under rational behavior."

In the continuum version, each player i's set of strategies is given by $S_i = [2,100]$. Let T_i be a subset of $S_i, i = 1,2$. The pair (T_1, T_2) is defined as *curb* (actually "tight curb" in Basu and Weibull [1991]) if T_i is the set of all best responses of player i to j's strategies in T_j, $i = 1,2$. In other words, the strategy s_1 belongs to T_1 if and only if there exists a strategy s_2 in T_2 such that player 1 cannot do better by unilaterally deviating from (s_1, s_2).[4]

A direct application of curb to the continuum version of the traveler's dilemma is not possible because there are no best responses in this game. However, here is a modified version of curb—I shall call it M-curb—which uses the *idea* of curb. Let (T_1, T_2) be called *$*$-curb* if T_1 and T_2 are nonempty and, for all s_2 in T_2 and all s_1, in S_1, there exists r_1 in T_1 such that player 1 does at least as well by responding to s_2 with r_1 instead of s_1, and likewise with players 1 and 2 interchanged.

(T_1, T_2) is *M-curb* if it is $*$-curb and *individually* minimal, that is, there does not exist M_1 which is a proper subset of T_1 or M_2 which is a proper subset of T_2 such that (M_1, T_2) is $*$-curb or (T_1, M_2) is $*$-curb.

It is easy to see that if (T_1, T_2) is M-curb then $\max[T_j]$ is either 2 or does not exist, for $j = 1$ or 2. Here is an example of an M-curb set: [(90,100), (90,100)]. Hence, if each player commits to play in the open interval (90,100), then no player has the incentive to deviate. While this is a resolution of the continuum version, this cannot be taken as a resolution of the paradox, because the heart of the paradox does not lie in the technical matter of whether players are allowed to use all real numbers or not; I will now turn to the integers version.

Possibility 2.—Though the traveler's dilemma is a normal-form game, it nevertheless can be thought of as having the

[2] For different perspectives on the standard backward-induction paradox, see, for instance, Frederic Schick (1983), Amartya Sen (1985), Michael Taylor (1987), Giacomo Bonanno (1991), Tilman Borgers and Larry Samuelson (1992), Martin Dufwenberg and Johann Linden (1993), and Martin Hollis and Sugden (1993).

[3] Confronted by a similar problem involving the iterated deletion of dominated strategies, Jacob Glazer and Rosenthal (1992) argue that players play cooperatively because they do not mind forgoing the small gains of noncooperative play. This may be so in reality, but my problem stems from the belief that even players who are scrupulous maximizers would play large numbers in a game like the traveler's dilemma.

[4] This is actually an imprecise definition of curb but it captures its essential idea.

"unreached-node problem." To see this, begin by (a) defining rational play in the usual way and then (b) assume that rationality is common knowledge.

Since, (2,2) is the only rationalizable outcome, it follows that that is what one should expect since rationalizability is the consequence of (a) and (b). Now suppose player 1 wants to decide how he would do if he rejected playing 2 and went instead for a larger number. It is not clear that this question is at all answerable. If it is true that (a) and (b) imply that player 1 will choose strategy 2, then a world where (a) and (b) are true and the player chooses some other strategy may not be conceivable, and so such introspective experiments may not be possible.

One possible line of attack that this suggests is to argue that the implicit assumptions, (a) and (b), which underlie so much of game theory, may by themselves be inconsistent. In Basu (1990) I showed that, in the context of games like centipede, the problem stemmed from assuming that rationality is common knowledge and that every game must have a solution. The method was to write down some properties of a solution, given that rationality is common knowledge, and to demonstrate that these properties cannot be together satisfied. However, there I made critical use of the extensive structure of the game. The traveler's dilemma challenges one to construct similar theorems without recourse to the sequence of play.

Possibility 3.—To end on an optimistic note, I shall now consider a more novel line of attack. Observe that in the traveler's dilemma there cannot exist a well-defined set of strategies, T_i, excepting the special case $T_i = \{2\}$, such that:

(i) a rational player may play any strategy in T_i and will never play anything outside it, and
(ii) such a T_i can be *deduced* from an examination of the game.

To see this suppose that T_1 and T_2 are such sets. Since player 2 is perfectly rational, he can deduce what the game-theorist can deduce. Hence, by (ii), he can deduce T_1. Let t be the largest number in T_1. Since player 1 will never play any number above t, it never pays for player 2 to play t. Hence T_1 and T_2 are not identical. But since this game is symmetric and T_1 and T_2 are deduced purely from examining the game, T_1 must be the same as T_2. This contradiction establishes that no such (T_1, T_2) exists.

Note that this whole exercise was for well-defined (i.e., the usual kind of) sets. Hence, there *may* exist ill-defined sets that would work. There seems to be some a priori ground for believing that there may be an escape route here. Harking back to an idea that was touched on earlier, suppose that player 1 believes that player 2 will play a large number. Then, if player 1 were simply deciding whether he himself should play a large number or not, it would be in his interest to play a large number. Thus (large, large) seems to be a kind of Nash equilibrium *in ill-defined categories*. The ill-definedness is important here because if the set of large numbers was a well-defined set, one knows from the above paragraph that this argument would break down.

I am here interpreting "large" in the sense of everyday language, which is different from the fuzzy-set-theoretic interpretation. The latter implies that the set of integers that are certainly not large is a well-defined or crisp set. The everyday use of the word "large" clearly does not conform to this. Once this is taken seriously, many objections concerning the idea of Nash equilibrium in ill-defined categories, which immediately come to mind, cease to be valid.

Consider a question like this: "if the other player is playing a large number, should I ever play a number 1 less than a large number?" Once one starts answering questions like this, the argument as to why (large, large) is a kind of Nash equilibrium will quickly break down. What I am arguing, however, is that the question like the one above is not permitted in this framework. Given the everyday use of the word "large," "a large number minus 1" is a meaningless term. All I am claiming here is that if a player is told that the other player will choose a large number and then asked whether he will choose a large number or not, he will say yes.

The use of imprecise categories does not mean forgoing rationality. What was argued in this subsection was that one way of holding on to the rationality assumption in the face of paradoxical games such as the traveler's dilemma may be to allow players to use ill-defined categories in doing their reasoning about how to choose in game-theoretic situations.[5]

REFERENCES

Bacharach, Michael. "Games with Concept-Sensitive Strategy Spaces." Mimeo, University of Oxford, 1991.

Basu, Kaushik. "On the Non-Existence of a Rationality Definition for Extensive Games." *International Journal of Game Theory*, 1990, *19*(1), pp. 33–44.

______. "Group Rationality, Utilitarianism and Escher's Waterfall." *Games and Economic Behavior*, 1994 (forthcoming).

Basu, Kaushik and Weibull, Jorgen. "Strategy Subsets Closed Under Rational Behaviour." *Economics Letters*, June 1991, *36*(2), pp. 141–46.

Bicchieri, Cristina. "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," in Wolfgang Balzer and Bert Hamminga, eds., *Philosophy of economics*. London: Kluwer, 1989, pp. 69–85.

Binmore, Ken. "Modeling Rational Players, Part I." *Economics and Philosophy*, October 1987, *3*(2), pp. 179–214.

Binmore, Ken and Brandenberger, Adam. "Common Knowledge and Game Theory," in K. Binmore, *Essays on the foundations of game theory*. Oxford: Blackwell, 1990, pp. 105–50.

Bonanno, Giacomo. "The Logic of Rational Play in Games of Perfect Information." *Economics and Philosophy*, April 1991, *7*(1), pp. 37–65.

Borgers, Tilman and Samuelson, Larry. "'Cautious' Utility Maximization and Iterated Weak Dominance." *International Journal of Game Theory*, 1992, *21*(1), pp. 13–27.

Dufwenberg, Martin and Linden, Johann. "Inconsistencies in Extensive Games." Working paper, Department of Economics, Uppsala University, 1993.

Glazer, Jacob and Rosenthal, Robert. "A Note on Abreu-Matsushima Mechanisms." *Econometrica*, November 1992, *60*(6), pp. 1435–38.

Hollis, Martin and Sugden, Robert. "Rationality in Action." *Mind*, January 1993, *102*(405), pp. 1–35.

Kreps, David; Milgrom, Paul; Roberts, John and Wilson, Robert. "Rational Cooperation in the Finitely-Repeated Prisoner's Dilemma." *Journal of Economic Theory*, August 1982, *27*(2), pp. 245–52.

Pettit, Philip and Sugden, Robert. "The Backward Induction Paradox." *Journal of Philosophy*, April 1989, *86*(4), pp. 169–82.

Reny, Phil. "Common Belief and the Theory of Games with Perfect Information." *Journal of Economic Theory*, April 1993, *59*(2), pp. 257–74.

Rosenthal, Robert. "Games of Perfect Information, Predatory Pricing and the Chain Store Paradox." *Journal of Economic Theory*, August 1981, *25*(1), pp. 92–100.

Schick, Frederic. *Having reasons*. Princeton, NJ: Princeton University Press, 1983.

Selten, Reinhart. "The Chain Store Paradox." *Theory and Decision*, April 1978, *9*(2), pp. 127–59.

Sen, Amartya. "Goals, Commitment and Identity." *Journal of Law, Economics, and Organization*, Fall 1985, *1*(2), pp. 341–55.

Taylor, Michael. *The possibility of cooperation*. Cambridge: Cambridge University Press, 1987.

[5] For an ingenious related analysis, see Michael Bacharach (1991).