The New Palgrave Dictionary of Economics and the Law, (ed. P. Newman), Macmillan, 1998

P

pain and suffering. *See* MEDICAL MALPRACTICE; PRODUCTS LIABILITY.

paradoxes of game theory. In recent years, game theory, as a method of analysis, has spread from its origins in economics to many other disciplines such as sociology, politics and, more recently, law (Baird, Gertner and Picker 1994). With the rise of this method have emerged some intriguing paradoxes which have both cast shadows on the foundations of rational behaviour models and strengthened the foundations by compelling us to reconsider and hone some of the standard axioms.

From the time of the Greeks, Zeno and Eubulides, in the fourth and fifth centuries BC, paradoxes, by appealing simultaneously to our senses of fun and philosophic wonder, have been a major instigator of scientific quests. This is quite evident in game theory, where paradoxes have typically arisen when the conclusions about human behaviour arrived at through the use of formal analysis have conflicted with our intuitive (and even reasoned) view of the matter. To the extent that people do not have identical intuitions, there can always be open questions about what constitutes a paradox. However, in game theory a set of problems that have arisen from the use of the 'backward induction argument' is widely accepted as deeply paradoxical.

Even though many of these paradoxes remain unsolved, they have enriched our understanding of strategic interactions between agents. Indeed some important works on entry-deterrence and collusion in industry – topics of interest to those dealing with antitrust legislation – have been inspired by the effort to grapple with the problem of backward induction in the Prisoner's Dilemma and related games (Selten 1978; Kreps, Milgrom, Roberts and Wilson 1982). Such analysis has also thrown light on financial economics and bubbles (Morris and Shin 1995) and related paradoxes have helped us pose questions of importance in ethics (Basu 1994b). Moreover, there are paradoxical results in game theory (e.g. Gale and Stewart 1953) which have opened up debates in mathematics: for instance, the possibility of a set theory without the axiom of choice.

THE BACKWARD-INDUCTION CONUNDRUM. An early encounter of economists with the problems of backward induction occurred when considering finitely-repeated plays of the Prisoner's Dilemma game (Luce and Raiffa 1957). Let us here diverge from that tradition and introduce the problem by analysing the less well-known game of Centipede due to Rosenthal (1981).

In the Centipede two players play alternately for up to 100 periods. It is easiest to think of it as a game in which player A has a parcel to start with. In period 1 person A can either keep the parcel (play K) or pass it to the other player (play P). If he chooses P then it is player B's move who can keep it or pass it. If she passes it, it is A's move again and he can keep it or pass it. The game is terminated as soon as someone chooses to keep the parcel or if it is passed 100 times. Each act of passing the parcel yields 2 dollars for each player. An act of keeping the parcel yields 3 dollars to the keeper. This game (with its hundred legs) is described in Figure 1.

For every pair of numbers in the above game tree the top number is A's payoff and the bottom number B's payoff. The game begins at node x_1 with A's move. If A chooses K, the game ends with A collecting a total of 3 and B collecting 0, as shown. If A passes, we reach node x_2, where B has to move. If B chooses K, the game ends and A gets a total of 2 and B a total of 5 (2 for the one pass and 3 for the keep). If B chooses P, it is A's move; if A chooses K the game ends with A collecting 7 and B collecting 4.

How will rational players play this game? *If* node x_{100} is reached, clearly it is rational for B to choose K instead of P (she gets \$201 instead of \$200). Since A can see this, at node x_{99} A will clearly choose K. That way he gets \$199. If he had chosen P, then (given B's anticipated move in the

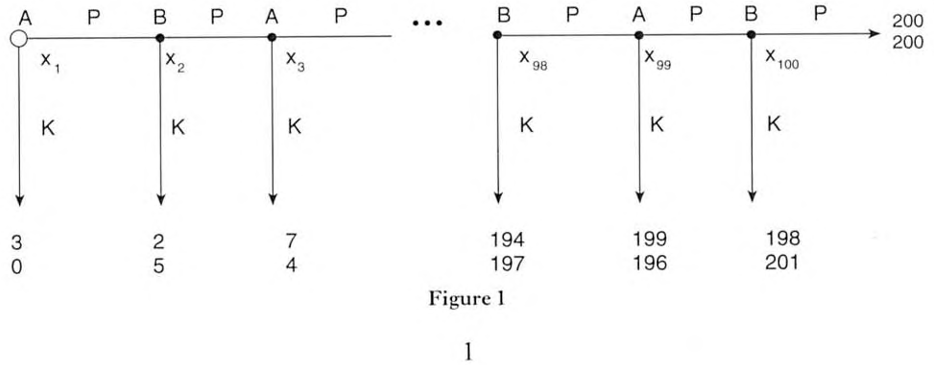

Figure 1

last period) he would get 198. It follows, by a similar reasoning, that if node x_{98} were reached, it would be rational for B to play K. And this argument unfolds inexorably backwards and takes us all the way to the prediction that in period 1 player A will choose K, collect \$3 and bring the game to an end. This is the so-called 'backward-induction' argument – an instrument of immense power (Aumann 1995) but also the cause of much philosophical dispute (Pettit and Sugden 1989).

The end result of the backward induction argument assaults our commonsense and intuition. Surely most reasonable people would play 'cooperatively' by choosing P, at least in the early games, expect the other player to do the same and earn considerably more than 3. But at the same time the backward-induction argument seems quite sound. It is this conflict which constitutes the paradox.

To the question why people may in reality play cooperatively in the early games, we can give a variety of answers. Players may be altruistic and give some weight to others' payoffs. A player may be rational but not know that the other player is rational or not know that the other player knows that he is rational. Incorporating such assumptions have enriched game theory and the study of law and economics of industry, but they do not solve the paradox. This is because there is a sense in us which suggests that, if players are ruthlessly selfish, know that they are rational, know that they know that they are rational and so on, *even then* they will *choose* to ignore the backward-induction argument and play P, at least in the early games, and expect their opponent to do the same. A large number of papers (e.g. Binmore 1987; Bicchieri 1989; Reny 1993) have considered problems of this kind and tried to explain how players will actually play such games. But open questions have continued to plague the field. Basu (1990) takes a different line and argues that games such as the Centipede are *unsolvable*. That is, *any* prediction of how such a game will be played will be inconsistent with some elementary axioms of rationality. It is important to appreciate that to say that a game has no solution is not the same as saying that 'anything can happen'. The latter is tantamount to saying that the solution consists of all possible outcomes. The theorem in Basu (1990), on the other hand, says that such a prediction is as wrong as the prediction that A will choose K in period 1.

Papers which have tried to solve this paradox have often exploited the extensive-form structure of the above game, that is, the fact that these games – Centipede or the Repeated Prisoner's Dilemma – are played over time. This is what allows players to 'throw surprises' by deviating from the path of backward induction (Binmore and Brandenburger 1990). Thus if in the Centipede, in the first move A chooses P, in effect A is giving a message to B that her backward-induction reasoning is demonstrably false. This can induce B to play cooperatively.

It is however possible to argue that the paradox runs deeper and can arise even in a single-shot game through a kind of *introspective* backward induction. This problem, with related discussions in Abreu and Matsushima (1992) and Glazer and Rosenthal (1992), is captured by the Traveller's Dilemma game (Basu 1994a; see also Zambrano 1996).

THE TRAVELLER'S DILEMMA. Two travellers returning home from a remote island where they bought identical antiques discover that the airline has managed to damage these. The airline manager, on the grounds that he has no way of confirming the price of these antiques, offers the travellers the following compensation scheme.

Each of the two travellers has to write down on a piece of paper the cost of the antique. This can be any integer 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 reasons the manager) and so each of these travellers will be paid n_1 (or n_2) units of money.

If traveller i writes a larger number than the other (i.e., $n_i > n_j$), then assume 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 traveller i the sum of $n_j - 2$ and pay j the sum of $n_j + 2$. Traveller i is paid 2 units less as penalty for lying and j is paid 2 units more as reward for honesty.

Given that each traveller 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? In a manner reminiscent of Keynes's elegant metaphor of the beauty contest (1936, chapter 12, section v), in this game the true value of the antique turns out to be irrelevant.

At first sight it appears that both players can get 100 by simply writing 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 (and even if the other player writes something other than 100, this player can never do worse by writing 99 instead of 100). So he should write 99. 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. 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. Indeed all standard solution concepts – Nash, strict Nash, rationalizability – predict the unique outcome (2, 2). Yet it seems unlikely that two individuals, no matter how rational they are, will play (2, 2). Of course, altruism, reciprocity and such elements of reality can explain why people will play large numbers. But the paradox occurs because it seems that they should and will play large numbers even if they decide to play selfishly.

What we need to grapple with is the fact that human beings are equipped with a higher order rationality which urges us, in contexts such as the Traveller's Dilemma or the Centipede, to reject formal reasoning and choose a large number; and assures us that the other player will do the same. But this is an argument that has proved hard to formalize and so we must move on, leaving this as an open-ended research problem.

KNOWLEDGE AND COMMON KNOWLEDGE. One assumption that is quite ubiquitous in game theory is that rationality is 'common knowledge' among the players. What this means is that the players are rational, each player knows that all players are rational, each player knows that all players know that all player are rational and so on. Game theorists and economists have usually been quite cavalier about this assumption, invoking it at will. The emergence of paradoxes of the kind discussed above have made some analysts wonder if the common knowledge assumption is not at the root of some of these paradoxes.

Indeed it is somewhat reminiscent of the way in which in early set theory it used to be implicitly assumed that there is a universal set which contains everything; and it was only with the appearance of paradoxes such as the celebrated 'Russell paradox' that it became clear that this seemingly innocuous assumption was a hornet's nest of inconsistencies.

Hence, starting with the work of the philosopher David Lewis (1969), there has now been much formal investigation into the algebra of knowledge and common knowledge (Fagin, Halpern, Moses and Vardi 1995). While these investigations have not really 'solved' the paradoxes of backward induction, they have both deepened our understanding of the relation between rationality and knowledge and drawn our attention to some new paradoxes, some of which will be discussed in the next section. But before going on to that it is useful to illustrate the somewhat mysterious nature of knowledge with Littlewood's (1953) story of the ladies in the train, with dirt on their faces, laughing at one another, each unaware that she has dirt on her face.

A more sterilized version of that story has it that a class has thirty red-haired students. Assume that no one can see the colour of his own hair but of course knows that others have red hair. This school has a rule that, if a person knows that he has red hair, he should not come to school. One morning the teacher announces in class that at least one student has red hair. At one level this gives no information and so should have no effect on any one's behaviour. But in this case after thirty (school) days no student returns to school.

To understand this first consider a class with only two students. After an announcement by the teacher that at least one person has red hair, in two days' time neither student will return to school. This is because when on the day following the announcement student *i* sees that *j* has come to school, *i* will realize that *i* must have red hair because, otherwise, hearing the teacher's announcement *j* would have realized that *j* has red hair and so *j* would not have come to school. Hence, by the second day each person knows that each person has red hair. To get to the class with 30 students we have to proceed by induction. Basically what happens is that the depth of knowledge (I know that you know that I . . .) keeps increasing with each passing day.

THE E-MAIL GAME. Rubinstein's (1989) celebrated example of the electronic mail game highlights a paradoxical nature of knowledge of different levels of depth and also demonstrates the sharply different implications of assuming common knowledge and assuming nearly common knowledge. This is related to the work of Gray (1978) and Halpern and Moses (1990).

Two friends, 1 and 2, plan to go to a bistro (B) or an amphitheatre (A) near 1's home. The amphitheatre is their preferred destination if and only if the weather is sunny (*s*). If it is rainy (*r*), they would rather go to the bistro.

If the weather is sunny, the game is G_s; and, if it is rainy, the game is G_r with payoffs as illustrated in Figure 2.

It is assumed that $p < \frac{1}{2}$ and $L > M > 1$. Note that in game G_s, for both players, A is a dominant strategy. That is, no matter what the other player does each person is better off going to the amphitheatre. In game G_r, on the other hand, (B, B) is the best outcome but choosing B is not a risk-free decision on the part of an individual, because if the other agent chooses A, then the individual will get $-L$.

Let us now assume that the weather condition is known only to player 1 who then communicates this to 2 by the following technology. If and only if it is rainy a message goes from 1's computer to 2's computer. From then onwards the two computers are programmed so that whenever a computer receives a message it automatically sends out a message (of acknowledgment). However, computers are not infallible so every time a message goes out there is a small probability, ϵ, that the message never reaches the other player's machine. So to sum up, if the weather is sunny, neither computer sends out a message because the process of sending messages never gets started. If the weather is rainy, (the machine of) player 1 sends at least one message with probability 1, he sends at least two messages with probability $(1-\epsilon)^2$ and so on, and player 2 sends at least one message with probability $1-\epsilon$, at least two messages with probability $(1-\epsilon)^3$ and so on. After the machines stop sending messages, each player checks how many messages were sent from his machine and chooses between going to A and B.

Suppose both machines send 10 messages. Then clearly it is rainy, both players know it is rainy and 1 knows that 2 knows that it is rainy; both players know that; both players know that both players know that and so on, up to ten times. The question is: After receiving those ten messages how will the players play? It can be shown that the only rational way to play is for both to choose A and earn zero!

The argument builds on induction by first considering the case when both players receive zero messages then

	A	B
A	*M*, *M*	1, $-L$
B	$-L$, 1	0, 0

G_s (Probability $1-p$)

	A	B
A	0, 0	1, $-L$
B	$-L$, 1	*M*, *M*

G_r (Probability p).

Figure 2

considering cases with one more message at a time. The interested reader is referred to Rubinstein (1989) for a proof. But one can see the essence of this argument by considering the same story but with M set equal to one. This is a less interesting game because A is now a dominant strategy also in G_r; but (B, B) is still the best outcome in game G_r and so we would expect both players to choose B if they know that it is rainy. It is easy to see that no matter how many messages go back and forth, it is rational for both players to play only A. It is true that (A, A) is a perfect equilibrium (in the sense of game theory), which (B, B) is not, but in this game that is not the reason why A is always played.

To see this, suppose both players' machines send zero messages. Then player 1 knows it is sunny and plays A. And player 2 thinks that either it is sunny or 1's message got lost. Hence, he believes there is a positive probability that 1 will play A. But then it is best for 2 to play A (remember we are now assuming $M = 1$).

Next consider the case in which 1's machine sends one message. This means that either the message from 1 fails to reach 2 or 2's message (i.e. 2's acknowledgement of 1's message) fails to reach 1, because if neither of these happened then 1's machine would have sent more than one message. It follows that 1 will know that either 2's machine sent zero messages or one message. If the former happens, we know from the above paragraph that 2 will play A. Since 1 knows this, 1 will play A.

Now consider 2's machine sends one message. Then 2 will know that either 1's machine sent one message or two messages. And we can continue to reason in this fashion.

If state r (rainy) were common knowledge, it is a Nash equilibrium for both to choose B and earn M each. But anything short of common knowledge, as we just saw, destroys this outcome. Both may know r, know that they know r and so on a hundred times, but the only rational play will be (A, A).

Once again, as with the backward induction paradox, one is left with the feeling that if I am in a situation where my machine sends out 100 messages, I will *choose* to ignore all this fine reasoning and play B and expect player 2 (who would have sent 99 or 100 messages) to do the same. But until this 'idea' is formalized the paradox must be treated as an unresolved one.

IMPERFECT RECALL. Another paradox dealing with knowledge and information, though of a very different genre from the E-mail game, is Piccione and Rubinstein's (1996) paradox of absent-mindedness. Ever since the 1950s it has been a staple assumption that players in a game have perfect recall – that is, at every stage of the game each player remembers what the player did and what the player knew in earlier stages. Piccione and Rubinstein have demonstrated that abandoning perfect recall is not just inconvenient but it may give rise to paradoxes.

Consider a one-player game in which a person sitting in a bar is contemplating driving back home. For that he has to take the second exit. Reaching home gives him a payoff of 4. The second-best option (payoff 1) is to take no exit and to reach a motel at the end of the road. The worst option is to take the first exit and reach a bad area (payoff 0). The trouble is, he is absent-minded and when he sees an exit he can never remember whether he has already gone past an exit or not. In other words, in the game described in Figure 3, he cannot tell the difference between nodes d_1 and d_2. Formally, d_1 and d_2 belong to the same information set.

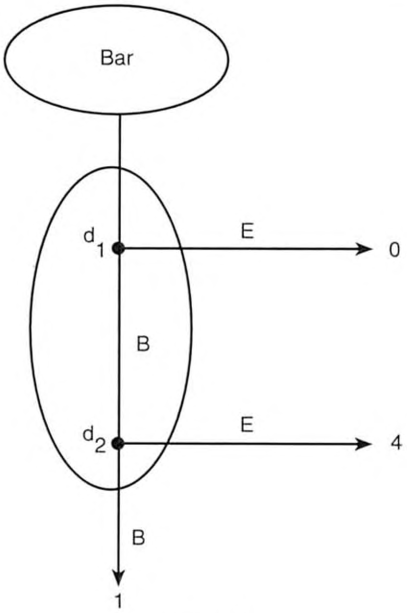

Figure 3

Given this handicap, he realizes in the bar that all he can decide is whether or not to take an exit when he sees it; and it is clear that his best strategy is not to take the exit. That way he will get a payoff of 1, instead of 0.

Now let us suppose that, having made that decision, he sets out. Soon he sees an exit. Given his own decision at the bar, he knows that this may be exit 1 or 2. Both are equally likely since he will go past both. So if he takes the exist his expected payoff is 2. That being greater than 1, he should take the exit. Nothing that he did not expect has happened since he was at the bar, but there seems to be reason for him to change his decision.

The paradox has generated a lot of controversy, which it is perhaps too early to assess properly. What the paradox seems to me to call into question is the status of information sets as primitives. If a person can remember his choice of strategy and has perfect powers of deduction this *may* be inconsistent with the assumption that the person is absent-minded. Also, once perfect recall is dropped as an assumption there may arise a case for dropping the assumption that information can be characterized as a *partition*.

To see this, return to the game in Figure 3 and suppose that the terminal node where the payoff is zero now gives a payoff of 5. In this modified game, his decision in the bar will clearly be to take the exit when he sees one. Hence, if he remembers this then at d_1 he will know that he is at d_1 and at d_2 he will 'know' (or think he knows) that he is at d_1. This implies a representation of knowledge which violates the standard axiom which asserts that, if a person knows an event, then that event must have occurred.

Paradoxical results such as these merely highlight the fact that we still have a great distance to go in understanding the relation between rational behaviour and knowledge.

KAUSHIK BASU

See also COMMON KNOWLEDGE; CONVENTIONS; GAME THEORY AND STATES OF THE WORLD; PRISONERS' DILEMMA.

Subject classification: 1d(i).

BIBLIOGRAPHY

Abreu, D. and Matsushima, H. 1992. Virtual implementation in iteratively undominated strategies: Complete information. *Econometrica* 60(5): 993–1008.

Aumann, R.J. 1995. Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8: 6–19.

Baird, D.G., Gertner, R.H. and Picker, R.C. (eds). 1994. *Game Theory and the Law*. Cambridge, MA: Harvard University Press.

Basu, K. 1990. On the nonexistence of a rationality definition for extensive games. *International Journal of Game Theory* 19(1): 33–44.

Basu, K. 1994a. The traveler's dilemma: Paradoxes of rationality in game theory. *American Economic Review* 84(2): 391–5.

Basu, K. 1994b. Group rationality, utilitarianism, and Escher's waterfall. *Games and Economic Behavior* 7: 1–9.

Bicchieri, C. 1989. Self-refuting theories of strategic interaction: a paradox of common knowledge. *Erkenntnis* 30: 69–85.

Binmore, K.G. 1987. Modeling rational players. *Economics and Philosophy* 3: 179–214.

Binmore, K.G. and Brandenburger, A. 1990. Common knowledge and game theory. In *Essays on the Foundations of Game Theory*, ed. K.G. Binmore, Oxford: Blackwell.

Fagin, R., Halpern, J.Y., Moses, Y. and Vardi, M.Y. 1995. *Reasoning about Knowledge*. Cambridge, MA.: MIT Press.

Gale, D. and Stewart, F.H. 1953. Infinite games with perfect information. In *Contributions to the Theory of Games II*, ed. H.W. Kuhn and A.W. Tucker, Princeton: Princeton University Press.

Glazer, J. and Rosenthal, R.W. 1992. A note on Abreu–Matsushima mechanisms. *Econometrica* 60(6): 1435–8.

Gray, J. 1978. *Notes on database operating systems.* In *Operating Systems: An Advanced Course*, ed. R. Bayer, R.M. Graham and G. Seegmuller, Berlin: Springer-Verlag.

Halpern, J.Y. and Moses, Y. 1990. Knowledge and common knowledge in a distributed environment. *Journal of ACM* 37(3): 549–87.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kreps, D., Milgrom, P., Roberts, J. and Wilson, C. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27: 253–79.

Lewis, D.K. 1969. *Convention*. Cambridge, MA: Harvard University Press.

Littlewood, J.E. 1953. *Mathematical Miscellany*. Ed. B. Bollobas, Cambridge: Cambridge University Press.

Luce, R.D. and Raiffa, H. 1957. *Games and Decisions*. New York: John Wiley.

Morris, S. and Shin, H.S. 1995. Informational events which trigger currency attacks. Mimeo: University of Pennsylvania.

Pettit, P. and Sugden, R. 1989. The backward induction paradox. *Journal of Philosophy* 86(4): 169–82.

Piccione, M. and Rubinstein, A. 1996. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, forthcoming.

Reny, P.J. 1993. Common belief and the theory of games with perfect information. *Journal of Economic Theory* 59: 257–74.

Rosenthal, R.W. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* 25(1): 92–101.

Rubinstein, A. 1989. The electronic mail game: Strategic behaviour under 'almost common knowledge'. *American Economic Review* 79: 385–91.

Selten, R. 1978. The chain store paradox. *Theory and Decision* 9(2): 127–59.

Zambrano, E. 1996. The algebra of inexact knowledge with an application to the game of Hermes. Paper presented at the ASSET conference, University of Alicante, 1996.