Colman, A. M., Krockow, E. M., Frosch, C. A., & Pulford, B. D. (2017). Rationality and backward induction in Centipede games. In N. Galbraith, E. Lucas, & D. E. Over (Eds.), *The thinking mind: A Festschrift for Ken Manktelow* (pp. 139–150). London: Routledge.

Rationality and Backward Induction in Centipede Games

Andrew M. Colman, Eva M. Krockow, Caren A. Frosch, and Briony D. Pulford
University of Leicester

Author Note

Andrew M. Colman, Eva M. Krockow, Caren Frosch, and Briony D. Pulford, School of Psychology, University of Leicester.

We are grateful to the Leicester Judgment and Decision Making Endowment Fund (Grant RM43G0176) for support in the preparation of this chapter.

Correspondence concerning this chapter should be addressed to Andrew M. Colman, School of Psychology, University of Leicester, Leicester LE1 7RH, United Kingdom. E-mail: amc@le.ac.uk

Among all the thousands of strategic games that have been discovered and investigated since the early development of game theory in the 1920s, especially after the publication of von Neumann and Morgenstern’s (1944) landmark book, *Games and Economic Behavior*, the Centipede game stands out as perhaps the most perplexing and paradoxical of them all. It was introduced by Rosenthal (1981) as an incidental comment (pp. 96–97) in a discussion of an entirely different game (the Chain-store game). The Centipede game was first named in print by Binmore (1987) after the passing resemblance of its game tree to a multi-legged insect, as can be seen in Figure 1, where Rosenthal’s original version of the game is depicted.

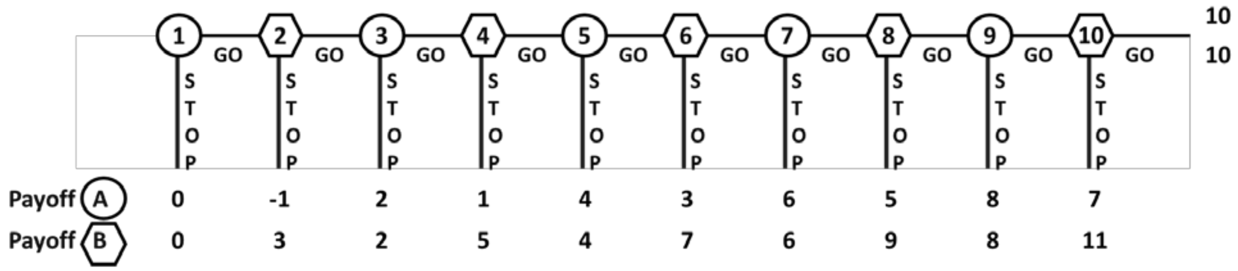

Figure 1. Game tree of Rosenthal’s (1981) original (as yet unnamed) Centipede game.

The rules of the game are simple. Starting at the left, Player A makes the initial move at the first numbered *decision node* by choosing either STOP or GO. Choosing STOP causes the game to end at that point (the label attached to this option gives a big clue), and choosing GO leads to the second numbered decision node, where Player B chooses between STOP and GO. Play continues in this fashion, with Players A and B taking turns choosing moves until one of them chooses STOP. If neither player chooses STOP at any of the 10 decision nodes, then the game ends naturally after the final node. The numbers in the *terminal nodes* at the feet of the Centipede and on its antenna protruding to the right are the payoffs to the players when the game ends, either because one of them chooses STOP or because the game reaches its natural end. Following the normal convention in game theory, the first payoff in each terminal node is for Player A and the second for Player B. Hence, if Player A chooses STOP at the first decision node, then each player receives a zero payoff; if Player A chooses GO at the first decision node and Player B chooses STOP at the second, then Player A loses 1 unit of payoff and Player B gains 3 units; and so on. Formally, the payoffs are measured in units of von Neumann–Morgenstern *utilities*, and these reflect the players' true preferences, incorporating all their tastes and predilections, selfish or altruistic motives, and so on, because they are defined in terms of the players' choices. But it is simpler to think of the payoffs as monetary units such as pounds sterling, euros, or dollars, and this is harmless enough for most purposes.

Rosenthal's original version in Figure 1 is a linear Centipede game, because the sum of payoffs to the player pair increases linearly across successive terminal nodes: the sums are 0, 2, 4, 6, . . . , 20, increasing by exactly two units from one terminal node to the next. A STOP move brings the game to an end without affecting the payoffs; a GO move alters the payoffs that the players have accumulated up to that point, invariably imposing a cost c (1 unit in this particular version of the game) on the player who chooses it and conferring a benefit b (3 units in this case) on the co-player. A GO move is altruistic inasmuch as it benefits another individual at some cost to the player who chooses it (a standard definition of altruism); but it is conventional to identify GO moves with cooperation and STOP moves with defection, because repeated GO moves benefit the player pair, whereas a STOP move provides a short-term selfish advantage to the individual player who chooses it. The number of decision nodes is arbitrary, and adding further legs to the Centipede does not affect any of its key strategic properties.

The fact that the payoffs shown in the terminal nodes increase as the game progresses is not because the cost and benefit parameters c and b change (they remain fixed) but merely because payoffs accumulate if both players keep choosing GO moves—a point that mathematically challenged reviewers and editors sometimes struggle to grasp. A linear Centipede game always has fixed c and b, with $c \leq b$, and it provides a simple model of repeated interactions between two individuals with alternating opportunities for reciprocal altruism and continual temptations for unilateral defection. An everyday example that we have frequently used is a reciprocally cooperative relationship between two university researchers who take turns reviewing and providing feedback on each other's manuscripts and grant applications, each reviewing task imposing a small cost on the reviewer but providing a larger benefit to the recipient. It is not difficult to see that everyday professional, economic, and social life involves numerous interactions with the general strategic structure of the linear Centipede game.

What, then, is so perplexing and paradoxical about this game? The answer emerges from a *backward induction* argument, so called because of its resemblance to mathematical induction, one of the standard techniques for proving theorems. The argument appears to establish that there should be no cooperation, and that a rational Player A should defect at the first decision node, yielding zero payoffs to both players. To reach this conclusion, the argument begins by assuming that the 10th decision node has been reached, where Player B faces the final decision in the game. Player B must choose between defecting and earning a payoff of 11 or cooperating and receiving a payoff of 10 when the game terminates naturally. To specify rational choice in this (or any) game, we usually base our deductions on standard game-theoretic *common knowledge and rationality* assumptions:

1. Rationality assumption: Both players are instrumentally rational in the sense that they have unlimited cognitive capacities and, whenever faced with a choice between two options with known payoffs, they invariably choose the one yielding the higher individual payoff to themselves;

2. Knowledge assumption: Both players know the rules of the game and the players' preferences, represented by the payoffs shown in the game tree;
3. Common knowledge assumption: Assumptions (1) and (2) are *common knowledge* in the sense that both players know them, both know that both know them, and so on ad infinitum (although only a finite number of iterations are required in a finite game, as we shall see).

Given these assumptions, at the 10th decision node, Player B will choose STOP. This follows from Assumptions 1 and 2, because STOP is the payoff-maximizing option (11 is better than 10), and Player B knows this without even considering what Player A knows. We can now deduce that Player A will defect at the immediately preceding ninth decision node, because the choice there is between defecting and earning a payoff of 8 or cooperating and earning a payoff of 7 when Player B defects on the next move. This follows from Assumptions 1 and 2, as before, plus one iteration of the common knowledge Assumption 3: Player A needs to know that B is instrumentally rational, otherwise A would not be certain that B would defect at the 10th decision node, if given the chance, and without this knowledge, A's choice at the 9th decision node would not be strictly determined. Following this line of reasoning further, it becomes clear that B will defect at the eighth decision node. This follows from all three common knowledge and rationality assumptions, and in this case B needs to know that A is rational and that A knows that B is rational, hence another iteration of the common knowledge Assumption 3 is required. Continuing in the same way, the backward induction argument unfolds, move by move, requiring one more iteration of common knowledge for each move, and it always leads to the conclusion that the rational move is to defect. Eventually, we arrive at the first decision node, where Player A will defect, because of Assumptions 1, 2, and nine iterations of Assumption 3. This outcome is the *subgame perfect Nash equilibrium* of the game, because it is, in effect, arrived at by the iterated elimination of weakly dominated strategies, and it is easy to prove that this ensures subgame perfection (a straightforward proof is provided by Osborne & Rubinstein, 1994, section 6.6.1, pp. 108–110). Game theorists acknowledge this as the uniquely rational outcome of this game.

The backward induction argument appears to establish that two rational players who understand the Centipede game in Figure 1 will earn precisely nothing from it, because Player A will defect at the first decision node. A similar conclusion applies even to exponential versions of the game, in which the sum of payoffs to the player pair is multiplied by a fixed amount at each terminal node rather than having a fixed amount added, as in a linear version. An extreme example of an exponential Centipede game, due to Aumann (1992),[1] is shown in Figure 2.

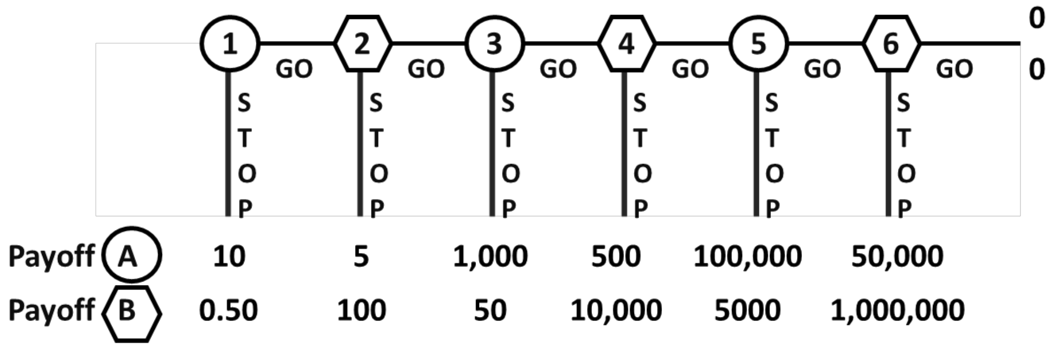

Figure 2. Game tree of Aumann's (1992) exponential Centipede game.

Note that Aumann's (1992) exponential Centipede game has zero payoffs in the final terminal node on the right; these are the payoffs that await the players if neither defects at any of the six decision nodes. This *zero-end* modification means that one of the players is virtually certain to defect at some point. The zero-end structure and the exponential form of the payoff function—payoffs mount up rapidly, increasing tenfold at each decision node, until the end—generate much more psychological pressure than Rosenthal's (1981) original version, but the logic of the game, and in particular the backward induction argument, remains the same despite these seemingly radical changes. Given the common knowledge and rationality assumptions, Player A will still defect at the first decision node, leading to the game's unique subgame perfect Nash equilibrium and yielding payoffs of 10 to A and 0.50 to B, although fabulous wealth is on offer if both players cooperate up to the sixth decision node.

Is the backward induction argument valid?

Can the backward induction argument possibly be valid? Instrumental rationality is defined as invariably choosing the option that maximizes one's own payoff, but backward induction leads to a solution that requires Player A to defect on the first move, and this actually *minimizes* Player A's payoff, quite literally in a standard Centipede game, and virtually in a zero-end version, where only the natural end is even worse. Game theorists acknowledge that this conclusion seems paradoxical, but the validity of the backward induction argument that underpins it has been endorsed by many authorities well qualified to make the call, including the Nobel laureate, mathematician, and game theorist Robert Aumann (1995, 1998). In one of his publications, Aumann (1992) proved that even the tiniest modicum of irrationality destroys the backward induction argument and justifies cooperation, to the players' mutual benefit. If the players are in fact perfectly rational, then a *belief* that there is a miniscule probability that one of them will deviate from rationality suffices to justify cooperation. Aumann even goes as far as to suggest that it is not necessary for either of the players to actually believe that one or both of them may be the tiniest bit irrational, provided that they tacitly agree to ignore the fact that both are fully rational:

> Most of us have experienced situations where some harmful fact is perfectly well known but is studiously overlooked by everybody. In this case, the harmful fact is the players' rationality(!). More precisely, the fact itself need not be harmful, but common knowledge of it would be. The above approach enables us to understand this phenomenon within the context of the theory. (Aumann, 1992, p. 226)

This seems to us an artificial and unsatisfactory workaround. It is inadequate because it leaves the absurd conclusion intact when common knowledge of rationality prevails without any ifs or buts; and it seems unnecessary if the absurd conclusion does not follow from the assumptions in the first place. We wish to argue, tentatively and with due deference to many of our distinguished game-theoretical colleagues, that the conclusion does not follow, and that the backward induction argument is fatally flawed.

According to the backward induction argument, at the second decision node (Figure 1 or Figure 2), Player B will defect, in the certain knowledge that a cooperative move would be followed by Player A defecting at the third decision node. This deduction is supposed to follow from the common knowledge and rationality assumptions. But, given those assumptions, what is B supposed to think when choosing a move at the second decision node? If B has the luxury of moving at all, A must have cooperated at the first decision node. According to the assumptions, B knows that A is instrumentally rational and that A knows that B is also instrumentally rational and would therefore defect at the second decision node, if given the chance, yielding a smaller payoff to A than the payoff if A defects at the first decision node. These two items of knowledge seem mutually contradictory: a rational Player A would not cooperate at the first decision node, because defection would pay better (in Figure 1, it is obvious that 0 is better than –1, and in Figure 2, that 10 is better than 5).

It seems to follow that Player B, contemplating a choice at the second decision node, is confronted with a situation that is literally impossible, given the common knowledge and rationality assumptions. Either Player A is not rational, or the payoffs are not as B believes them to be, or perhaps not as A believes them to be, or one of the players has misunderstood the rules of the game—something, at least, must have gone wrong. Furthermore, whatever it is that has gone wrong makes it impossible for B to predict how A would respond to a cooperative move. A Player A who is willing to cooperate at the first decision node may cooperate again at the third; there is simply no way of knowing, because the theory has collapsed and the principles supposedly determining A's actions have been invalidated. Player B simply has no rational basis for choosing a move at the second decision node, and the same applies a fortiori at later nodes.

The backward induction argument begins with a consideration of B's move at the final decision node; but to have arrived at that point, several moves must have occurred that are inconsistent with the common knowledge and rationality assumptions. In both Figure 1 and Figure 2, Player B still has a valid reason to defect, because defection pays better than cooperation, and the common knowledge assumption is not required at that point. But, in Figure 1, at the previous (penultimate) decision node, Player A cannot know that defection pays better than cooperation, because B's response is indeterminate. At this penultimate decision node, A has no basis for rational choice, because the common knowledge assumption bears the load of the whole argument, and Player A knows that it is not true, because the game could not have progressed to that point if it were true. Nothing can validly be deduced about A's move, and the backward induction argument therefore collapses. The same applies to Player B at Node 4 in Figure 2.

Aumann was not unaware of this problem, of course; rather, he carefully sidestepped it by redefining rationality according to what follows and burying the past in a manner resembling feigned dense amnesia:

> Rationality of a player at a vertex *v* is defined in terms of what happens at vertices after *v*, his payoff *if v* is reached. The idea is that when programming his automaton at *v*, the player does so as if *v* will be reached—even when he knows that it will not be! Each choice must be rational "in its own right"—a player may not rely on what happened previously to get him "off the hook." (Aumann, 1995, p. 12)

This is the crux of our disagreement. In Figure 1, at the ninth decision node, Player A is well and truly "off the hook" because, without ignoring the past entirely, A has no way of knowing what will happen after a cooperative move, and therefore no way of deciding whether STOP or GO is the payoff-maximizing move. The same dilemma confronts Player B at the fourth decision node in Figure 2. Centipede is a game of perfect information, and there is nothing in the common knowledge and rationality assumptions, nor in any common-sense interpretation of rationality, to justify a pretence by a player that what is known to have happened previously did not happen. In game theory, players are usually assumed to know the specification of the game and everything that can validly be deduced from any moves that have been made. Aumann's (1995) suggestion seems to us an inadequate and, above all, unjustified solution to the crucial problem that undermines the backward induction argument.

Experimental evidence

Whatever the rights or wrongs of the backward induction argument, human decision makers certainly do not follow its prescriptions in incentivized experimental Centipede games. Experimental players are much more cooperative, and they therefore earn much better payoffs. McKelvey and Palfrey (1992) were the first to investigate decision making in six-node exponential Centipede games with non-zero payoffs in the terminal node at the end; they found that only 0.7% of all games ended at the first decision node, and 1.4% of the games even continued until the natural end. McKelvey and Palfrey reported some learning effects—more frequent use of competitive defecting strategies—with increasing experience in repetitions of the game, but these effects were comparatively small and did not result in equilibrium play. Similar results were found in a replication study reported by Kawagoe and Takizawa (2012) and in Nagel and Tang's (1998) investigation of learning in 12-node exponential Centipede games. Even higher levels of cooperation were reported for Centipede games with linearly increasing payoff functions, like Figure 1 (Bornstein, Kugler, & Ziegelmeyer, 2004; Gerber & Wichardt, 2010).

These studies provide overwhelming evidence against the use of backward induction reasoning in the Centipede game, inspiring Levitt, List and Sadoff (2011) to investigate the move choices of expert chess players (including several Grandmasters), who would arguably have greater iterated reasoning skills than the average undergraduate experimental participant. Interestingly, the chess players' choices were remarkably similar to those of previous student participants and adhered to the game-theoretical solution in only 3.9% of all games. Furthermore, not a single chess player with perfect backward induction reasoning ability (as measured with the Race to 100 game, a strictly competitive game also requiring backward induction reasoning for its solution) stopped the Centipede game at the first decision node. These findings suggest that cooperation in the Centipede game cannot be accounted for merely in terms of limited cognitive abilities. Instead, there is experimental evidence that other-regarding preferences such as concerns for the payoffs to the player pair influence the players' strategies (Pulford, Colman, Lawrence, & Krockow, 2015; Pulford, Krockow, Colman & Lawrence, 2015).

Backward induction in the history of game theory

The first formal theorem in the history of game theory was (arguably) based on backward induction. Long before von Neumann (1928) proved his famous minimax theorem, marking the official birthday of game theory, the German mathematician, Ernst Zermelo (1913) proved a theorem about strictly competitive games of *perfect information*—games in which each player knows all moves that have been made previously. He proved that chess (or any other such game) is *strictly determined*, in the sense that there is either a guaranteed winning strategy for one of the players or guaranteed drawing strategies for both. The proof does not provide any method for actually finding the winning or drawing strategies—in the case of chess it is still not known whether White (or conceivably Black) has a winning strategy or whether both players have drawing strategies, although the latter seems most likely. Most authorities (e.g., Binmore, 1992, p. 32; Fudenberg & Tirole, 1991, p. 91) interpret Zermelo's method of proof as a form of backward induction, although some (e.g., Schwalbe & Walker, 2001) disagree. Zermelo certainly began his proof by considering terminal positions—positions from which one of the players can force a win or a draw, and he proceeded roughly as follows. There must be only a finite set of such positions, because (in chess) there are only 64 squares and, at most, 32 pieces and pawns.

Assume that in all the positions that can lead immediately to those terminal positions the player whose turn it is invariably makes a best move (which must exist), and in all positions immediately preceding those, players also choose best moves. It must be possible in principle, though not in practice, to roll back through the unimaginably large game tree, specifying best moves in every position that could arise. This procedure would, in effect, construct a fully specified best-move strategy for each player, and bringing the two together would necessarily lead either to a win for White, a win for Black, or a draw.

Von Neumann and Morgenstern and Morgenstern (1944, chap. 15, pp. 112–128) surprisingly omitted to cite Zermelo's (1913) classic proof, but they included a detailed and formal discussion of backward induction in strictly competitive (two-player, zero-sum) games. Their presentation is long and abstract, and some of their mathematical notation obsolete, so that their chapter on backward induction is now largely inaccessible even to professional mathematicians.

		II	
		C	*D*
I	*C*	3, 3	0, 5
	D	5, 0	1, 1

Figure 3. The Prisoner's Dilemma game with conventional payoffs.

The locus classicus of backward induction is the discussion in the standard textbook of game theory by Luce and Raiffa (1957) of finitely repeated or *finite-horizon* Prisoner's Dilemma games. We shall outline their argument using the now conventional version of the game shown in Figure 3. Player I chooses a strategy corresponding to a row, either *C* (cooperate) or *D* (defect), Player II independently chooses a strategy corresponding to a column (*C* or *D*), and the numbers in the cell where the two strategy choices intersect are the resulting payoffs to Players I and II respectively for that outcome of the game. From Player I's point of view, *D* is a dominant strategy, because it yields a better payoff than *C* irrespective of Player II's strategy choice (5 rather than 3 if Player II chooses *C*; and 1 rather than 0 if Player II chooses *D*); and *D* is similarly a dominant strategy for Player II. This establishes that, for a single play of this game, defection is the only rational strategy for both players, despite the fact that both are better off if both cooperate. However, in an *infinite-horizon* version, or if the game is repeated an *indefinite* number of times between the same players, then it is no longer the case that defection is uniquely rational, because there are reasons for cooperating, including signaling to the co-player a conditional willingness to cooperate by playing Tit for Tat ("I'll cooperate if you will"). However, these reasons exist only when there are further rounds to be played, because they necessarily involve making sacrifices in the short run in the hope of larger payoffs later on.

What Luce and Raiffa (1957, section 5.5, pp. 97–102) showed was that, in a finite-horizon Prisoner's Dilemma game, if both players know in advance that exactly 100 rounds (for example) are to be played, then the only rational strategy is to defect on every round. The proof begins with the last round. Here there is no reason for either player to cooperate, because there are no subsequent rounds to be taken into consideration, hence the last round is, in effect, a one-shot Prisoner's Dilemma game, and because defection is a dominant strategy for both players, the only rational choice is to defect. Therefore, on the penultimate (99th) round, there is no reason to cooperate, because the outcome of the last round is predetermined and cannot be influenced, hence the only rational choice for both players is, once again, to defect. The argument unfolds, round by round, all the way back to the beginning, where the only rational choice for both players on the first round is to defect. This appears to prove that, if the common knowledge and rationality assumptions hold, then rational players will defect on every round of a finite-horizon Prisoner's Dilemma game, but it is not vulnerable to the objection that we raised to this argument in the Centipede game.

The unexpected hanging

The unexpected hanging is a well-known paradox based on backward induction, and it bears a superficial resemblance to the Centipede game. Sentencing a man on a Friday, a judge says: “You will be hanged one afternoon next week, beginning on Sunday, but you will not know the day of your execution until the morning of that day.” The prisoner reasons as follows. “I can't be hanged on Saturday, the last possible day, because by Friday night I would know that my day of execution has to be Saturday. Furthermore, I can't be hanged on Friday because, with Saturday ruled out, by Thursday night I would know that my day of execution must be Friday. Continuing in the same vein, I can exclude all the remaining days of the week. I can't be hanged on any day next week, and therefore the judge's sentence can't be carried out at all. Whoopee!”

This paradox is sometimes called the surprise examination paradox, because it can be framed as a warning by a teacher about an examination to be held the following week. It was discovered by the Swedish mathematician Lennart Ekbom in 1943 or 1944, during the Second World War, when the Swedish broadcasting system announced a civil defence drill for the following week and said that no one would know in advance on which day it would take place. It was first discussed in print by O'Connor (1948), who interpreted the judge's sentence as a classic self-defeating prophecy, and for that reason considered the problem “rather frivolous”. It was immediately pointed out by others, however, that if the hanging were to take place on Monday, for example, then it would indeed be unexpected, and hence that the judge's sentence is not a self-defeating prophecy after all: it does not predict something that cannot occur.

Many philosophers have commented on the paradox, and it has generated by far the largest body of literature of any epistemic paradox. In spite of all this attention, it is clearly a falsidical paradox, and not a veridical paradox (with a true conclusion) or an antinomy (a genuine logical contradiction). The prisoner argues that there is a contradiction between the judge's statements (a) that the sentence will be carried out the following week, and (b) that the prisoner will not know the day of his execution until the morning of that day, and that either (a) or (b) must therefore be false. He then infers invalidly that (a) rather than (b) is false, and therefore that the hanging cannot take place on any day of the following week. He provides no reason why it could not be (b) that is false: he could know the day of his execution before the morning of that day. Furthermore, he assumes that if he is not hanged by Friday, then he must be hanged on Saturday, and elsewhere he (inconsistently) assumes that if he is not hanged by Friday, then Saturday will have been "ruled out"; but he cannot have it both ways.

The philosopher Quine (1953) was the first to expose the prisoner's fallacious reasoning, with the following refutation. The prisoner claims that he cannot be hanged on Saturday because he would know his day of execution by Friday night; but this is false, because if he were to be hanged on Saturday, then by his own reasoning, it would indeed come as a surprise, and the same argument extends backwards to the earlier days of the week. Quine pointed out that if the judge had told the prisoner that he would be hanged *the following day* but would not know the day of his execution until the morning of that day, then it would be similarly incorrect of the prisoner to infer that he could not be hanged at all, because he could, in fact, be hanged the following day without knowing in advance that he would be hanged, because of the incoherence of the judge's sentence.

The unexpected hanging certainly involves backward induction reasoning, but its resemblance to the Centipede game is superficial. The way backward induction is applied to the Centipede game may be wrong, as we have claimed, but it is certainly not wrong for the reason that the prisoner's argument in the unexpected hanging is wrong. In particular, backward induction in the Centipede game does not reach a conclusion that could be refuted by events on the ground, as the prisoner's reasoning does.

Conclusions

Backward induction is not objectionable in itself, but it can be applied fallaciously. The way it is used in the unexpected hanging paradox seems clearly fallacious, and we have suggested that its usual application in the Centipede game may also be questionable, though for quite different reasons. If we are right, then there do not appear to be any compelling arguments against cooperation in the Centipede game, and indeed the experimental evidence confirms that intelligent human decision makers, including people capable of backward induction reasoning, almost invariably cooperate, at least for a few moves. There may be nothing strictly paradoxical about this game, although it does pose difficult strategic problems to players.

Endnotes

[1] Aumann (1992, p. 220) attributed this version to Megiddo (1986), but Megiddo's game does not have zeros in the final terminal node—its most original feature. One is reminded of the Viennese-born Jewish violinist Fritz Kreisler (1875–1962), who frequently performed compositions that he attributed to Vivaldi and others although he had, in fact, composed them himself. We believe that Aumann introduced the zero-end Centipede game himself in 1988, at a workshop in Haifa organized by the Israeli economist Joseph Greenberg.

References

Aumann (1992). Irrationality in game theory. In Dasgupta, P., Gale, D., Hart, O., & Maskin, E. (Eds.), *Economic analysis of markets and games* (pp. 214–227). Cambridge, MA: MIT Press.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior, 8*, 6–19. doi: 10.1016/S0899-8256(05)80015-6

Aumann, R. J. (1998). On the Centipede game. *Games and Economic Behavior, 23,* 97–105. doi: 10.1006/game.1997.0605

Binmore, K. (1987). Modeling rational players: Part 1. *Economics and Philosophy 3,* 179–214. doi:10.1017/S0266267100002893

Binmore, K. (1992). *Fun and games: A text on game theory.* Lexington: D. C. Heath.

Bornstein, G., Kugler, T., & Ziegelmeyer, A. (2004). Individual and group decisions in the centipede game: Are groups more "rational" players? *Journal of Experimental Social Psychology, 40,* 599–605. doi: 10.1016/j.jesp.2003.11.003

Fudenberg, D., & Tirole J. (1991). *Game theory.* Cambridge, MA: MIT Press.

Gerber, A., & Wichardt, P. C. (2010). Iterated reasoning and welfare-enhancing instruments in the Centipede game. *Journal of Economic Behavior & Organization, 74,* 123–136. doi: 10.1016/j.jebo.2009.12.004

Kawagoe, T., & Takizawa, H. (2012). Level-*k* analysis of experimental Centipede games. *Journal of Economic Behavior & Organization, 82,* 548–566. doi:10.1016/j.jebo.2012.03.010

Levitt, S. D., List, J. A., & Sadoff, S. E. (2011). Checkmate: Exploring backward induction among chess players. *American Economic Review, 101,* 975–990. doi: 10.1257/aer.101.2.975

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey.* NewYork: Wiley.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the Centipede game. *Econometrica, 60,* 803–836. doi: 10.2307/2951567

Megiddo, N. (1986). *Remarks on bounded rationality*. Research Report 5270 (54310). Yorktown Heights, NY: IBM Research Division.

Nagel, R., & Tang, F. (1998). Experimental results on the Centipede game in normal form: An investigation on learning. *Journal of Mathematical Psychology, 42,* 356–84. doi: 10.1006/jmps.1998.1225

O’Connor, D. J. (1948). Pragmatic paradoxes. *Mind, 57,* 358–359. doi:10.1093/mind/LVII.227.358

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, MA: MIT Press.

Pulford, B. D., Colman, A. M., Lawrence, C. L., & Krockow, E. M. (in press). Reasons for cooperating in repeated interactions: Social value orientations, fuzzy traces, reciprocity, and activity bias. *Decision.* doi: 10.1037/dec0000057

Pulford, B. D., Krockow, E. M., Colman, A. M., & Lawrence, C. L. (2016). Social value induction and cooperation in the Centipede game. *PLOS ONE, 11*(3), 1–21. e0152352. doi: 10.1371/journal.pone.0152352

Quine, W. V. O. (1953). On a so-called paradox. *Mind, 62,* 65–66. doi:10.1093/mind/LXII.245.65

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and chain-store paradox. *Journal of Economic Theory, 25*, 92–100. doi: 10.1016/0022-0531(81)90018-1

Schwalbe, U., & Walker, P. (2001). Zermelo and the early history of game theory. *Games and Economic Behavior, 34,* 123–137. doi:10.1006/game.2000.0794

von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100, 295–320. doi:10.1007/BF01448847

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Zermelo, E. (1913). Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. *Proceedings of the Fifth International Congress of Mathematicians, Cambridge, 2,* 501–504.