The New Riddle of Backward Induction

Johan E. Gustafsson and Wlodek Rabinowicz

Draft: December 7, 2023 at 11:24 a.m.

ABSTRACT. In the Centipede game, the standard backward-induction argument recommends the first player to immediately terminate the game. This is puzzling, since the players would be better off if they let the game continue for several rounds. Another part of this old riddle questions assumptions behind the standard argument. This argument implausibly assumes that the players at all nodes of the game, even those that aren't reachable by rational play, would act rationally and retain trust in the future rationality of all players. A more plausible, weak form of backward induction merely makes assumptions about what the players would believe at nodes that are reachable without anyone making irrational choices, and in particular assumes that trust in rationality of the players would be retained at such nodes. These weak assumptions suffice to prove that the first player in the Centipede would be irrational if she let the game continue. But, given a plausible story about what the second player would expect after being confronted with the first player's irrational move, that irrational move would predictably give the first player a better pay-off than the terminating move she is rationally required to make. If rational behaviour consists in the maximization of expected pay-off, we seem to have arrived at a contradiction. This is our new riddle of backward induction. We tentatively suggest a solution and draw an analogy between this new riddle and Gaifman's Irrational-Man paradox.

Backward induction is a method — seemingly, a compelling one — of solving sequential games (and sequential choice problems) by predicting what would be chosen at later choice nodes and then taking those predictions into account at earlier choice nodes. There is, however, an old riddle of backward induction. In a game like the Centipede, the standard backward induction argument recommends the first player to immediately terminate the game. This is puzzling, since the players would be much better off if they continued the game for several rounds.

Another, more deep going, part of the old riddle challenges the driving assumptions behind the standard argument. This argument assumes that the players, at all choice nodes of the game (including those that cannot be reached by rational play), would act rationally and retain their trust in the future rationality of all players. This is highly implausible. Why suppose that a player is bound to act rationally, if they acted irrationally in

the past?

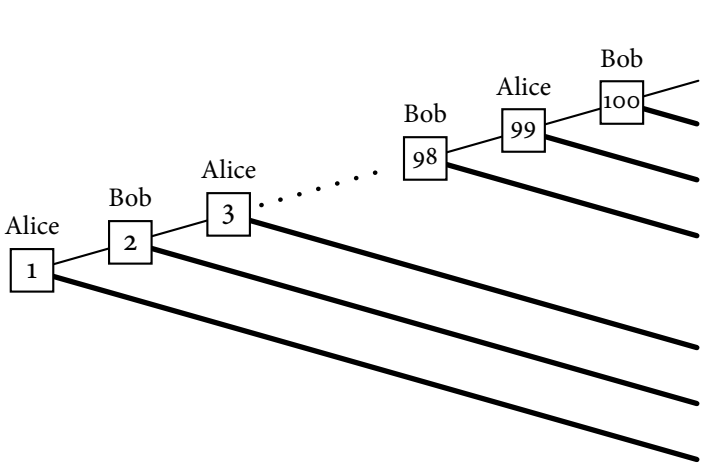
But there is a more plausible form of backward induction. That form of backward-induction reasoning merely makes assumptions about what the players would believe at nodes that are reachable without anyone making irrational choices, and in particular assumes that trust in rationality of the players would be retained at such nodes. These weak assumptions suffice to prove that the first player in the Centipede would be irrational if she let the game continue. Nevertheless, we will show that, given a plausible story about what the second player would expect when confronted with the first player's irrational move, that irrational move would predictably give the first player a better pay-off than the terminating move she is rationally required to make. Thus, if rationality consists in the maximization of expected pay-off (and we reject rational dilemmas), we seem to have arrive at a contradiction. This is our new riddle of induction.

We are going to suggest that this riddle has a solution, but that solution incurs a considerable cost: It requires that we give up the highly compelling idea that an action is irrational if one of its alternatives would predictably lead to an outcome that the agent prefers. And, in Appendix A, we draw an analogy between the new riddle of backward induction and Haim Gaifman's Irrational-Man paradox.

1. The old riddle

In the Centipede game we are going to consider, two players — call them Alice and Bob — take turns deciding whether to terminate the game. If the game is terminated at node n and n is odd, Alice (who moves at that node) gets a pay-off of $n + 1$ and Bob gets a pay-off of n. If the game is terminated at node n and n is even, Alice gets a pay-off of $n - 1$ and Bob (who moves at that node) gets a pay-off of $n+2$. Before the final round, the pay-offs are structured so that, if you defect (that is, terminate the game), you get a larger pay-off than if you cooperate (that is, let the game go on) and the other player defects at their next turn but a smaller pay-off than if you cooperate and the other player also cooperates at their next turn. And, if it's your turn to move at the final round, you get a larger pay-off if you defect than if you cooperate. At this final round, cooperation means making a move that benefits the other player at your own expense: The pay-offs in the final round coincide with the pay-offs the other player would have caused by defection in the next round if the game had one more round. Consider, as an illustration, the one-hundred-round version

of this game:[1]



	ALICE	BOB
	102	101
	99	102
	100	99
	97	100
	⋮	⋮
	4	3
	1	4
	2	1

Here, the boxes represent choice nodes where the player listed above the box makes a move — either letting the game continue (going up) or terminating it (going down). The table on the right lists the players' pay-offs in each outcome.

There is a standard backward-induction argument that each player is rationally required to go down at each choice node. At node 100, Bob would go down since that gives him a higher pay-off. Taking this into account at node 99, Alice would go down since going up would (given Bob's predicted choice at node 100) give her a pay-off of 99 whereas going down would give her a pay-off of 100. Taking this into account at node 98, Bob would go down since going up would (given Alice's predicted choice at node 99) give him a pay-off of 99 whereas going down would give him a pay-off of 100. And so on until we reach node 1, where Alice would go down since going up would (given Bob's predicted choice at node 2) give her a pay-off of 1 whereas going down would give her a pay-off of 2. (In the diagram above, the recommended moves are marked by the thicker lines.)

The recommendation to go down at node 1, however, seems paradoxical given that both players would be much better off if they started off cooperating (that is, going up) at a significant number of nodes. This is the old riddle of backward induction.[2]

[1] Rosenthal 1981, p. 96.

[2] Selten 1978, pp. 136–8 and Pettit and Sugden 1989, pp. 169–71.

Or rather, a part of it. The other part, which goes deeper, has to do with the assumptions underlying the standard backward-induction argument in favour of going down at the initial node. This argument assumes that the players at all nodes of the game would act rationally and retain trust in their own future rationality and the future rationality of the other players. But, if the conclusion of the argument is correct, then backward induction codifies rational behaviour in sequential games. This leads to a paradox: At a node that can only be reached by moves that contravene the recommendations of backward induction, the player whose turn it is to move has evidence that the players who moved at the previous nodes behaved irrationally (since they chose in violation of the recommendations of backward induction). But then, when confronted with such evidence, it would be epistemically irrational of the player to retain their trust in those other players' future rationality. Furthermore, if that player was one of those who made some such irrational moves at the preceding nodes, then this past irrational behaviour might negatively influence their current disposition to behave rationally. It may therefore be questioned whether the player would act rationally at the node under consideration. All this undermines the assumptions of rationality and trust in rationality on which the standard backward-induction argument has been relying in the first place.[3]

2. Getting by with weaker assumptions

A more plausible, weak form of backward induction assumes the following:

> *Trust in Rationality* If node *n* is reachable without anyone making irrational moves, then the player at the immediately preceding node would believe that the player at node *n* would not make an irrational move at node *n*.[4]

[3] Binmore 1987, pp. 196–200, Bicchieri 1988, pp. 145–7, Reny 1988, pp. 364–5, and Pettit and Sugden 1989, p. 172.

[4] For Trust in Rationality to be reasonable, we need to presuppose that the players do not mistakenly believe that some past moves in the game have been irrational when in fact they have not. Otherwise, it would be difficult to explain why they trust that the player who moves next won't make an irrational move. This presupposition is potentially controversial, but it could be justified if we suppose that the players' initial beliefs aren't excessively opinionated — that they start with beliefs that do not preclude any game development in which no one makes irrational moves. As a result, their initial trust in

Belief in Trust in Rationality At each node that can be reached without anyone making irrational moves, its player believes in Trust in Rationality.

Introspection At each node that can be reached without anyone making irrational moves, its player knows which of the available moves at that node are irrational and which are not.

Logical Competence At each node that can be reached without anyone making irrational moves, its player believes in what logically follows from what that player believes.

By irrational choices (or moves) we mean, here and in what follows, choices (moves) that are rationally prohibited. Correspondingly, a choice (move) is rational if and only if it is rationally permitted, while a player is rational if and only if her choices (moves) are rational, or at least not irrational.

Surprisingly, as will be proved below, this weak form of backward induction is actually sufficient to prove that, in the Centipede, it is rationally required to go down at node 1. This is so, since the game is *BI-terminating* — that is, each move that is prescribed by the standard form of backward induction terminates the game. For such games, weak assumptions of this kind about rationality and trust in rationality suffice to defend the backward-induction solution.[5] And Centipede games of any length are BI-terminating.[6]

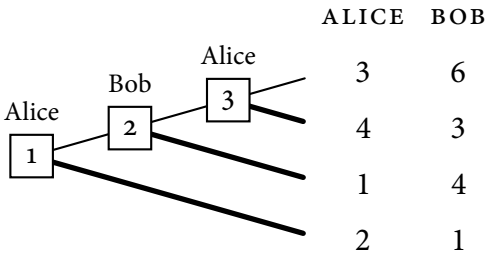
Rather than staying with the one-hundred-round version, we will show that Alice is rationally required to go down using the (more manageable) three-round version of the Centipede, but the argument can be extended to Centipede games of any length:

---

the players' rationality won't be undermined as long as no one acts irrationally. We are indebted to Robert Sugden, and to Robert Stalnaker, for alerting us to this issue.

[5] See Rabinowicz 1998.

[6] For the proof for Centipedes of any length, we need two further assumptions. See Appendix B.



Assume, for proof by contradiction, that node 3 can be reached without any irrational moves. Then — given Trust in Rationality — Bob at node 2 believes that Alice wouldn't make an irrational move at node 3. Since going up at node 3 gives Alice a lower pay-off than going down, it would be irrational for Alice to go up at node 3. Accordingly, Bob believes that Alice would go down at node 3, which means he believes that going up at node 2 would give him a lower pay-off than going down at that node (he would get a pay-off 3 rather than 4). Hence it's irrational for Bob to go up at node 2, which contradicts our assumption that node 3 can be reached without irrational moves. Thus node 3 cannot be reached without irrational moves.

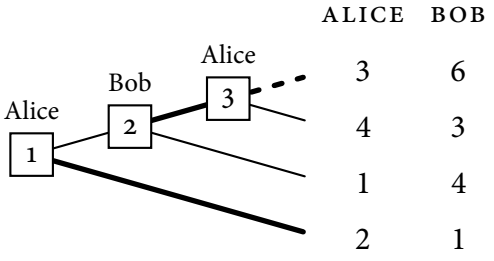
This conclusion entails that, if going up at node 1 is not irrational, it would be irrational for Bob to go up at node 2. Alice believes at node 1 that the premise used to derive this result holds: by Belief in Trust in Rationality, she believes in Trust in Rationality. Thus, by Logical Competence, Alice believes at node 1 that, if going up at node 1 is not irrational, then going up at node 2 is irrational.

Suppose now, for proof by contradiction, that going up at node 1 is not irrational. By Introspection, Alice at node 1 believes this. And, by Trust in Rationality, Alice believes that Bob would not make an irrational move at node 2. Since she also believes that going up at node 2 is irrational if going up at node 1 is not irrational, she believes at node 1, by Logical Competence, that Bob would go down at node 2. But then she believes at node 1 that going up at node 1 would give her a lower pay-off than going down. Hence it's irrational for her to go up at node 1, which contradicts our assumption. It follows that going up at node 1 is irrational.[7]

[7] This argument for defection in the Centipede is similar to the one in Broome and Rabinowicz 1999. See also Rabinowicz 1998, pp. 108–9 and Aumann 1998, p. 103. The sequential (extended) form of the game we study is crucial. Cubitt and Sugden (2014, pp. 295–6) argue that in a non-sequential (normal) form of this game (that is, one in which the players at the outset make a one-off choice between strategies instead of choos-

3. A new riddle

Assume now that the players (correctly) believe that *if* Alice (contrary to what they expect) were to make an irrational move at node 1, she would also make an irrational move at node 3 if the game were to reach that far. That is, Alice and Bob both believe that Alice would go up at node 3, even though going down would guarantee her a better pay-off. And assume that they are right in this belief and that Alice is aware of this belief on Bob's part. Note that these assumptions are consistent with the ones underlying the weak form of backward induction. So our earlier proof that it is rationally required to go down at node 1 still applies. We mark Alice's irrational move at node 3 (the move expected by both Alice and Bob if Alice were to reach that node) with a dashed line.



	ALICE	BOB
	3	6
	4	3
	1	4
	2	1

At node 2, Bob (who is trusted to be rational at that node, given that he hasn't previously made any irrational move) notes, with surprise, that Alice has made an irrational move at node 1. This leads Bob to (correctly) believe that Alice would (irrationally) go up at node 3. Taking this prediction into account at node 2, Bob sees that going up would give him a pay-off of 6 whereas going down would give him a pay-off of 4. Therefore, being rational, Bob would go up at node 2. Note that for Bob to be rational to go up at the second node, he need not be certain that Alice would then go up at the third node. It is enough if the probability of her

ing moves at the consecutive nodes) going down at node 1 can neither be shown to be rationally permitted nor to to be irrational. The key difference is this: In the sequential form, when Bob makes his choice at node 2 after Alice has gone up at node 1, he rules out her going down at node 1 but might well consider it possible, or even not unlikely, that she will also go up at node 3. While in the non-sequential form, when Bob makes his choice at the outset of the game, he might not rule out that Alice's strategy involves going up at node 1, but he certainly rules out that it also involves going up at node 3. At the outset of the game, when players make their strategy choices, Bob's belief in Alice's rationality is not undermined.

going up exceeds 1/3. This suffices for Bob's expected pay-off from going up to be larger than what he would get if he went down. (We could also change Bob's pay-off in the uppermost outcome from 6 to an arbitrarily large number — so that Bob would only need an arbitrarily small credence that Alice will go up at node 3 to make it rationally required for him to go up at node 2.[8])

As we have already shown, Alice is rationally required to go down at node 1. But note that she can predict that, if she were to go up at node 1, she would end up with a pay-off of 3 (because Bob would in such case go up at node 2 and she would then do the same at node 3). Whereas, if she were to go down at node 1, she would end up with a pay-off of just 2. Hence we have the paradoxical result that *it is rationally required to go down at node 1 even though going up would predictably give the player a higher pay-off*.[9]

Note that this paradox is not an argument that it is rationally permitted (or required) to go up at node 1. If going up at node 1 were permitted, it would no longer yield a higher pay-off than going down at that node.[10] Above, we have shown that weak assumptions about rationality

[8] But, even with this modification, one might wonder whether it is psychologically realistic of Bob to expect that Alice might act irrationally at the last node just because she started off the game with an irrational move. Perhaps not, but in our example we do not aspire to psychological realism. For our purposes, it is enough if the example is consistent with the weak assumptions about rationality and trust in rationality that we made in the preceding section.

[9] For a single-agent version of this paradox, consider an agent with cyclical preferences (or other non-standard preferences) who faces a BI-terminating money pump, such as the Upfront Money Pump (see Gustafsson and Rabinowicz 2020, p. 583). In such money pumps, the agent is rationally required by the weak form of backward induction to pay an exploiter to go away rather than to face a series of trades. But the agent may believe (consistently with the assumptions of that underlay this weak form of backward-induction reasoning) that if they (irrationally) did not pay the exploiter to go away they would also (rationally or irrationally) turn down the later trades. And then, believing so in advance, they would prefer the outcome of not paying the exploiter to the outcome of paying him, even though it is the latter that is rationally required. But, unlike the Centipede version of the paradox, the single-person one does not significantly challenge Dominance (which will be introduced in the next section), since an alternative resolution to the single-person version is that the agent's cyclical preferences (or other non-standard preferences) are irrational.

[10] Our earlier proof rules out that going up at node 1 is rationally permitted. It may seem that there is a conflict between the two results, but there isn't. In the earlier proof, we showed that

(i) If it is rationally permitted for Alice to go up at node 1, then her pay-off would

and trust in rationality suffice to prove that Alice would be irrational if she were to go up at the first node. Rationality requires her to go down. But, given additional and plausible assumptions about what a player would (correctly) believe if they were confronted with an irrational move by the other player, Alice's irrational move would predictably give her a better pay-off than the move she is rationally required to make.[11] But how is it possible? If rational behaviour consists in the maximization of expected pay-off, then going down at node 1 is also irrational. So, given that there are no rational prohibition dilemmas (that is, nodes where all options are irrational), we seem to have arrived at a contradiction.[12] This is our new riddle of backward induction.[13]

There is a notable difference between the new riddle and the old one. The old riddle was predicated on the assumption that backward induction codifies rationality at all choice nodes, which has led to the paradoxical

---

be lower if she were to go up than if she were to go down.

Now, we have have shown that

(ii) If it is not rationally permitted for Alice to go up at node 1, then her pay-off would higher if she were to go up than if she were to go down.

Claims (i) and (ii) are consistent. While these results are compatible, there may, as we shall see later, still be a conflict between the new upshot and a principle we have implicitly relied on in our earlier proof.

[11] Note that this is different from the less perplexing cases where it is rationally required to intentionally make oneself irrational. See Schelling 1960, p. 18 and Parfit 1984, pp. 12–13. In fact, even the weak form of backward induction rules out that a rationally permitted choice at a node at which its agent hasn't yet made any irrational choices in the past could predictably lead to the agent choosing irrationally at some future node. To allow for this, we would have to weaken the assumptions of backward induction even further. In Schelling's and Parfit's cases, you have both opportunity and reason to make yourself irrational in the future with the help of an irrationality drug. Making use of the drug leads to a preferred outcome even on the supposition that it is rationally permitted to do so. Contrast this with Alice's move up in the first node of the Centipede. Its preferred predicted outcome essentially depends on it being irrational.

[12] For prohibition dilemmas, see Vallentyne 1989, p. 302.

[13] This may seem similar to the 'Why ain'cha rich?'-objection to causal decision theory's two-box recommendation in Newcomb's problem. See Nozick 1969, p. 115, Gibbard and Harper 1978, p. 153, and Lewis 1981b. But there is an important difference between the two objections. In the Newcomb Problem, one-boxers become millionaires, as opposed to two-boxers. But there is no suggestion that a two-boxer would become a millionaire if she took just one box. (If she is a two-boxer, there is no million in that box.) While in the case we consider, we have argued that Alice (who is rational and will therefore go down at node 1) would end up with a higher pay-off if she chose to go up at that node.

conclusion that if both players were to act irrationally in a number of initial rounds, they would both get better pay-offs than if they were rational. According to the new riddle, the irrational move of the first player would give her a better pay-off than the rational move (that is, defection). But, if both the first and the second payer were to act irrationally, then the first player would go up and the second would go down (that is, defect). This combination of irrational moves would not give any of the two players a better pay-off than that player's rational move. Thus the two riddles are different.

Moreover, and more importantly, the old riddle was posed as a problem for the standard backward-induction argument — an argument that relied on the strong assumption that every player at every node would act rationally and have trust in the future rationality of all players. On this assumption, Alice would go down at node 3, and Bob, expecting this, would go down at node 2. Consequently, Alice's irrational move at node 1 would predictably give her a lower pay-off than her rational move at that node. The new riddle of backward induction can only arise if the implausibly strong assumptions about rationality and trust in rationality are weakened, as it was done in section 2.

4. Suggesting a solution

Is there a contradiction in the new riddle? Does one part of it presuppose what the other part denies? Maybe and maybe not. The proof we presented earlier, to the conclusion that it is rational for Alice to go down at the first node, implicitly relied on the following principle:[14]

> *Dominance* At a node where player *S* has finitely many available moves, a move is irrational if its expected outcome (as determined by *S*'s beliefs and credences) is less preferred by *S* than that of some other available move.[15]

But then we presented an argument to the effect that, if Alice at the first node were to choose the irrational option (that is, if she went up), her

[14] Our proof also assumed, implicitly, that it is common knowledge between the players that Dominance holds.

[15] Davidson et al. 1955, p. 145. We have added the restriction to nodes with a finite number of options to avoid cases where all options are dominated. See Nozick 1963, p. 89.

pay-off would predictably be higher than if she were to act rationally and went down. This argument thus undermines the very principle (that is, Dominance) on which the earlier proof was based. How can going down be rationally permitted if it would lead to a lower pay-off?

There may, however, be a way to avoid this inconsistency. As the reader can check, the earlier proof would still go through if, instead of Dominance, it relied on the following alternative principle:

> *Conditioned Dominance* At a node where player *S* has finitely many available options, an option x is irrational if its expected outcome on the hypothetical assumption that x is not irrational is less preferred by *S* than that of some other available option y on the assumption that x is not irrational.

In interpreting how Conditioned Dominance is supposed to be understood, it is important to clarify how we think of the hypothetical assumption that an option is not irrational. In hypothetically assuming this, we do not envisage any modification in the factual circumstances of the case that are grounds of the option's rationality status. The potential modification that is being envisaged only concerns the rationality status itself of the option in question and the expected effects that the recognition of its rationality is going to have on the beliefs and behaviour of the players at subsequent choice nodes.[16]

Option x (such as Alice's going up at the first node) on the hypothetical assumption that x is not irrational may be shown to give the agent a lower pay-off than its alternative y (Alice's going down) on the hypothetical assumption that y is not irrational, and yet, at the same time, it can be argued that x would give Alice a predictably higher pay-off than y if we in this argument start from the recognition that x is irrational.[17] This

[16] Another way to spell this out would instead be in subjunctive terms, as follows: We assess each option by what its predicted outcome would be if that option were not irrational (by a local rational miracle in case the option actually is irrational). These local rationality miracles are analogous to Lewis's (1979, p. 468; 1981a, p. 117) local divergence miracles with respect to the laws of nature — which, according to Lewis, need to be posited, given determinism, to account for subjunctive conditionals with false antecedents. So, if an option is irrational, we imagine that the principles of rationality would be just like they actually are except that they would not prohibit the option and that each rational agent would know this.

[17] As the reader can check, if — as we assume — Alice's irrational move at node 1 would lead Bob to expect that Alice would also act irrationally at node 3, Contingent

allows us to avoid the apparent inconsistency between the earlier proof and the argument that followed.

But can we give up the intuitively compelling Dominance? Its hold on us is hard to shatter. It is therefore not obvious that the contradiction we have pointed to can be avoided. We leave this question to the reader.

Appendices

A. The Irrational Man

The Irrational Man is a classical rationality paradox, due to Haim Gaifman.[18] It was slightly modified by Robert C. Koons, and then additionally modified by Vann McGee.[19] Especially McGee's version exhibits striking similarities to our new riddle of backward induction. It goes like this: You have a choice between *A*, an empty box, and *B*, a box containing $100. You are promised, by a reliable promisor, that if you choose *A and* this choice is irrational (but not otherwise), you will receive $1000.

It might seem that there is a contradiction here — if we assume that at least one option in the case must be rationally permitted. My choice of *A* cannot be rationally permitted, for if it were, it wouldn't be rewarded and thus would give me a lower pay-off than if I had chosen *B*. On the other hand, if my choice of *A* is irrational, then (given our assumption above) it must be rationally permitted to choose *B* instead. But if the choice of *A* is irrational, it would have been amply rewarded. Surely, it can't be rationally permitted to choose *B* if I would receive more had I chosen *A*?[20]

Dominance still suffices to establish that it would be rationally forbidden for Bob to go down at node 2 and thus that he could be expected to go up — thereby making Alice's irrational move at node 1 advantageous to her.

[18] Gaifman 1983, p. 150. Gaifman credits G. Schwartz with first suggesting this paradox.

[19] Koons 1992, pp. 17–19 and McGee 1993, p. 665. For yet another version, see Gaifman 1999, p. 120.

[20] One might also consider another rationality paradox that in some ways is simpler and yet also exhibits this similarity with our riddle. Thus consider the following irrationality bet, which is analogous to Alice's choice at node 1 in our riddle:

(I) If accepting bet (I) is irrational, you win 1 util; otherwise, you lose 1 util.

Arguably, it is rational to reject this bet and irrational to accept it. And yet, accepting it, while irrational, would give you a more preferred outcome (1 util instead of 0). If rationality consists in maximization of expected pay-off, bet (I) reveals a self-referential circularity: Its pay-off, and thus its rationality status, depends on its rationality status. It

If our solution of the new riddle of backward induction is applied to this paradox, there is no longer any incoherence. The argument above, which purports to establish a contradiction, rests in its last step on Dominance. Option *B* is supposed to be rationally prohibited if its alternative, *A*, has a preferred predicted outcome. If Dominance is given up and replaced by Conditioned Dominance, the contradiction disappears. Given the latter principle, *B* would be rationally prohibited if its predicted outcome were less preferred than that of *A* on the assumption that *A* is not irrational. But, on that assumption, taking *A* would not be rewarded and thus the outcome of that option would be less preferred than that of *B*. Therefore, there is no inconsistency in the suggestion that taking *B* is rational in the case at hand, and that it would be irrational to take *A*, even though the latter option has a preferred predicted outcome.[21] This is analogous to what we have encountered in the Centipede: It is rational for Alice to go down at the first node and it would be irrational for her to go

---

may not be obvious how circularity is present in the game-theoretic riddle, but note that, if rationality consists in maximization of expected pay-off, then the rationality of going up at node 1 depends (in part) on the expected pay-off of that move for its player and this pay-off in turn depends on the irrationality of that move. Thus our game-theoretical paradox is in some ways related to other self-referentially circular paradoxes such as *the Liar*, 'This sentence is false.' (See Cicero *Acad.* 2.95–6; 2006, pp. 55–6 and Mates 1981, pp. 15–40.) Gaifman (1983, p. 150) likewise notes the similarity between the Liar and the Irrational Man.

[21] But why can't we allow that *A* also is rationally permitted in this case, along with *B*? Conditioned Dominance excludes this. On the assumption that *A* is not irrational, *A* leads to a less preferred outcome than *B* on the assumption that *B* is rationally permitted. Therefore, Conditioned Dominance implies that *A* is irrational.

up, even though the latter move has a preferred predicted outcome.[22]

This analogy, however, should not hinder us from recognizing important differences between the game-theoretic riddle and the Irrational Man. In our discussion of the latter, we have assumed that at least one of the options, *A* or *B*, has to be rationally permitted. There was no need to make the corresponding assumption in the game-theoretic case. (We did, however, assume that not all options are rationally prohibited.)

But what if we abstain from this assumption in the Irrational Man? Then the paradox might be solvable in other ways as well. It might then be suggested, for example, that both options that are at the agent's disposal are irrational (if this is at all possible), even though only one, option *A*, would be rewarded with a thousand-dollar bonus. It might not be clear which solution of the rationality paradox is most satisfactory. We should also note that it isn't obvious that every option must be either rational (that is, rationally permitted) or irrational (rationally prohibited). Allowing for options that are neither opens up further possibilities that need to be considered.[23]

[22] We can also construct an analogous paradox for consequentialism — or, specifically, for the following principle:

> *Consequentialist Dominance* If (at a node with a finite number of options) the consequences of option *x* are worse than the consequences of some other available option, then *x* is morally wrong.

Consider *the Immoral Man*, where option *A* brings about 1 unit of value and option *B* brings about 2 units of value but, if you choose *A and* this choice is morally wrong (but not otherwise), a demon will bring about 2 additional units of value. Assuming that at least one option is morally right, Consequentialist Dominance leads to a contradiction no matter whether *A* is or is not morally wrong. We can avoid this paradox by instead accepting the following analogue of Conditioned Dominance:

> *Conditioned Consequentialist Dominance* If (at a node with a finite number of options) the consequences of option *x* on the hypothetical assumption that *x* is not morally wrong are worse than the consequences of some other available option *y* on the hypothetical assumption that *y* is not morally wrong, then *x* is morally wrong.

On the hypothetical assumption that *A* is not morally wrong, the consequences of *A* would be worse than those of *B* on the assumption that *B* is not morally wrong. Thus, by Conditioned Consequentialist Dominance, *A* is morally wrong. This is so despite the fact that, due to the intervention of the demon, this wrong option would lead to better consequences than the morally right *B*. (Conditioned Consequentialist Dominance is intended as a moral principle rather than a principle for choice under moral uncertainty. Using the principle for the latter would lead to very implausible recommendations.)

[23] See Gaifman 1983, p. 152.

Another important difference between the Irrational Man and our game-theoretic riddle is that the former paradox arises from letting the rational status of options be part of the description of their outcomes, whereas the latter does not. This makes it clearer, for our game-theoretic riddle, that the underlying decision problem is consistent.[24]

B. Centepedes of any length

To extend the argument for defection at the initial node to Centipedes of an arbitrary length n, we make the following additional assumptions:

> *Trust in Past Rationality* At each node that is reachable without anyone making any irrational moves, the player at that node believes that the node was reached in that way.[25]
>
> *Common Belief* It is a common belief among the players at nodes that are reachable without irrational loves that Trust in Rationality, Trust in Past rationality and Logical Competence hold.[26]

We are going to prove, by induction, that for no node i ($1 < i < n$) does it hold that i can be reached without irrational moves.

Base step: Assume, for proof by contradiction, that node n can be reached without any irrational moves. Then, by Trust in Rationality, the player at node $n - 1$ believes that the player at node n wouldn't make an irrational move at node n. Since going up at node n gives the player at that node a lower pay-off than going down, it would be irrational to go up at node n. Accordingly, the player at node $n - 1$ believes that the player at node n would go down at node n, which means the player at node $n - 1$

[24] We sidestep, for instance, the inconsistency objections considered in Gaifman 1999, p. 122.

[25] Just as Trust in Rationality, we can justify Trust in Past Rationality if we suppose that the players' initial beliefs aren't excessively opinionated — that they start with beliefs that do not preclude any game development in which no one makes irrational moves. As a result, their initial trust in the players' rationality won't be undermined as long as no one acts irrationally.

[26] In other words, (1) at each node that is reachable without irrational moves, its player believes at that node that Trust in Rationality, Trust in Past rationality and Logical Competence hold; (2) at each node that is reachable without irrational moves, its player believes at that node that (1) holds; (3) at each node that is reachable without irrational moves, its player believes at that node that (2) holds; and so on.

believes that going up at node $n-1$ would give them a lower pay-off than going down at that node. Hence it is irrational to go up at node $n-1$, which contradicts our assumption that node n can be reached without irrational moves. Thus we have proved that node n cannot be reached without irrational moves.

Inductive step: Suppose that, using as premises our assumptions (Trust in Rationality, Trust in Past Rationality, Introspection, Logical Competence, and Common Belief), we have proved that node $i+1$ (where $1 < i < n$) cannot be reached without irrational moves. We now want to prove that the same applies to node i. Suppose, for proof by contradiction, that node i can be reached without irrational moves. By Introspection and Trust in Past Rationality, (i) the player at node $i-1$ believes that node i can be reached without irrational moves. (Proof: If node i is reachable without irrational moves, then the same applies to node $i-1$. Hence, by Trust in Past Rationality, (a) the player at node $i-1$ believes that node $i-1$ has been reached without irrational moves. And since node i is supposed to be reachable without irrational moves, the move from node $i-1$ to node i is not irrational. Which implies, by Introspection, that (b) the player at node $i-1$ believes that the move from node $i-1$ to node i is not irrational. Given (a) and (b), by Logical Competence, player at node $i-1$ believes that node i is reachable without irrational moves.) By Common Belief, a player at a node that is reachable without irrational moves believes Trust in Rationality, Trust in Past rationality, Introspection, Logical Competence, and indeed Common belief itself. This implies that the player at node $i-1$ believes the premises of the proof that node $i+1$ cannot be reached without irrational moves. Hence, by Logical Competence, the player at that node $i-1$ believes that, (ii) if node i can be reached without irrational moves, then going up at node i is irrational. Given (i) and (ii), by Logical Competence, the player at node $i-1$ believes that going up at node i is irrational. And, given (i) and Trust in Rationality, the player at node $i-1$ believes that the player at node i would not make an irrational move at node i. Hence, by Logical Competence, the player at node $i-1$ believes that the player at node i would go down at that node. But then the player at node $i-1$ believes that going up at that node would give them a lower pay-off than going down. Hence it is irrational to go up at node $i-1$. Thus node i cannot be reached irrational moves — which contradicts our assumption. Hence node i cannot be reached without irrational moves.

The above inductive proof establishes, for all nodes i (where $1 < i <$

n) that node i cannot be reached without irrational moves. This holds, in particular, for $i = 1$, which mean that going up at node 1 would be irrational. This concludes our proof.

We wish to thank John Broome, Caspar Hare, Robert C. Stalnaker, Robert Sugden, Christian Tarsney, and the audience at the philosophy colloquium at MIT on November 3, 2023 for valuable comments.

References

Aumann, Robert J. (1998) 'On the Centipede Game', *Games and Economic Behavior* 23 (1): 97–105.

Bicchieri, Cristina (1988) 'Strategic Behavior and Counterfactuals', *Synthese* 76 (1): 135–169.

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

Broome, John and Wlodek Rabinowicz (1999) 'Backwards Induction in the Centipede Game', *Analysis* 59 (4): 237–242.

Cicero (2006) *On Academic Scepticism*, Indianapolis: Hackett.

Cubitt, Robert P. and Robert Sugden (2014) 'Common Reasoning in Games: A Lewisian Analysis of Common Knowledge of Rationality', *Economics and Philosophy* 30 (3): 285–329.

Davidson, Donald, J. C. C. McKinsey, and Patrick Suppes (1955) 'Outlines of a Formal Theory of Value, I', *Philosophy of Science* 22 (2): 140–160.

Gaifman, Haim (1983) 'Paradoxes of Infinity and Self-Applications, I', *Erkenntnis* 20 (2): 131–155.

— (1999) 'Self-Reference and the Acyclicity of Rational Choice', *Annals of Pure and Applied Logic* 96 (1–3): 117–140.

Gibbard, Allan and William L. Harper (1978) 'Counterfactuals and Two Kinds of Expected Utility', in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory*, vol. I, pp. 125–162, Dordrecht: Reidel.

Gustafsson, Johan E. and Wlodek Rabinowicz (2020) 'A Simpler, More Compelling Money Pump with Foresight', *The Journal of Philosophy* 117 (10): 578–589.

Koons, Robert C. (1992) *Paradoxes of Belief and Strategic Rationality*, Cambridge: Cambridge University Press.

Lewis, David (1979) 'Counterfactual Dependence and Time's Arrow', *Noûs* 13 (4): 455–476.

— (1981a) 'Are We Free to Break the Laws?', *Theoria* 47 (3): 113–121.
— (1981b) "Why Ain'cha Rich?", *Noûs* 15 (3): 377–380.
Mates, Benson (1981) *Skeptical Essays*, Chicago: University of Chicago Press.
McGee, Vann (1993) 'Review of Robert C. Koons, *Paradoxes of Belief and Strategic Rationality*', *Mind* 102 (408): 665–668.
Nozick, Robert (1963) *The Normative Theory of Individual Choice*, Ph.D. thesis, Princeton University.
— (1969) 'Newcomb's Problem and Two Principles of Choice', in Nicholas Rescher, ed., *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday*, pp. 114–146, Dordrecht: Reidel.
Parfit, Derek (1984) *Reasons and Persons*, Oxford: Clarendon Press.
Pettit, Philip and Robert Sugden (1989) 'The Backward Induction Paradox', *The Journal of Philosophy* 86 (4): 169–182.
Rabinowicz, Wlodek (1998) 'Grappling with the Centipede: Defence of Backward Induction for BI-Terminating Games', *Economics and Philosophy* 14 (1): 95–126.
Reny, Philip J. (1988) 'Common Knowledge and Games with Perfect Information', *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1988 (2): 363–369.
Rosenthal, Robert W. (1981) 'Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox', *Journal of Economic Theory* 25 (1): 92–100.
Schelling, Thomas S. (1960) *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.
Selten, Reinhard (1978) 'The Chain Store Paradox', *Theory and Decision* 9 (2): 127–159.
Vallentyne, Peter (1989) 'Two Types of Moral Dilemmas', *Erkenntnis* 30 (3): 301–318.