The New Riddle of Backward Induction

Johan E. Gustafsson and Wlodek Rabinowicz*

Draft
April 22, 2024

ABSTRACT. In the Centipede game, the standard backward-induction argument recommends the first player to immediately terminate the game. This argument implausibly assumes that the players at all choice 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. 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 in determining what should be chosen 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 can't be reached by rational play), would act rationally and retain their trust in

* We would be grateful for any thoughts or comments on this paper, which can be sent to *johan.eric.gustafsson@gmail.com*.

the future rationality of all players. This is highly implausible. Why suppose that a player is bound to act rationally and be trusted to do so, if they acted irrationally in the past?

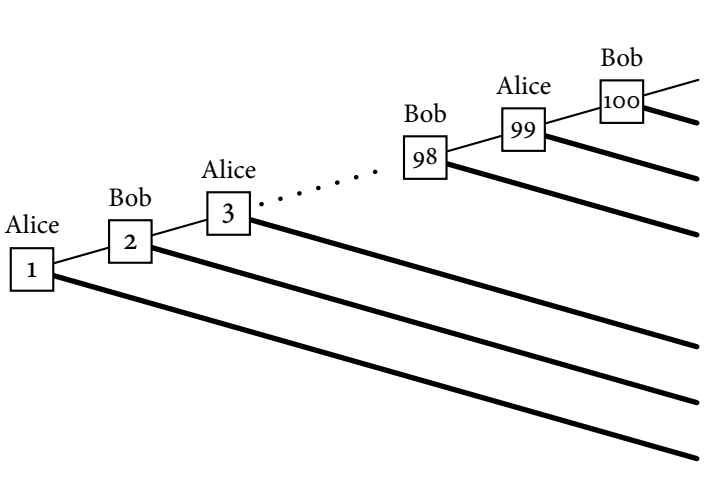
But there is a more plausible form of backward induction. That form of backward-induction reasoning merely makes assumptions about what the players would believe at nodes that are reachable without anyone making irrational choices and, in particular, assumes that trust in rationality of the players would 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. If rationality consists in the maximization of expected pay-off, it would follow that the terminating move at the first node must also be irrational. But if we reject the possibility that all moves that are available to a player at a given node can be irrational (that is, rationally prohibited), we seem to have arrived 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 an appendix (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 $n + 2$. If it's your turn to move prior to the final round, your pay-off from defection (that is, from the terminating move) is larger than if you cooperate (that is, if you let the game go on) but the next player defects, but it is smaller than if you cooperate and the next player also cooperates. 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

[1] Rosenthal 1981, p. 96.

cooperating (that is, going up) at a significant number of nodes. This is one part of the old riddle of backward induction.[2]

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 choice 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]

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

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

[4] In order 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

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

Introspection At each node that can be reached without anyone making irrational moves, it holds that, if a move available to the player at that node is not irrational, then the player at that node believes it is not irrational.

Logical Competence At each node that can be reached without anyone making irrational moves, the player at that node 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.

As will be proved below, this weak form of backward induction is actually sufficient to establish 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 suffice to defend the backward-induction solution.[5] And Centipede games of any length are BI-terminating.[6]

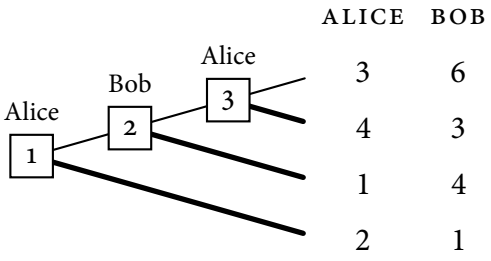
Rather than staying with the one-hundred-round version, we will show that Alice is rationally required to go down using the (more manageable) three-round version of the Centipede, but the argument can be extended to Centipede games of any length:

---

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 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] But, for a proof that holds for BI-terminating games of any length, the assumptions we have introduced above wouldn't suffice; we are going to 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. By Belief in Trust in Rationality, Alice believes at node 1 the premise used to derive this result: she believes that Trust in Rationality holds. Hence, 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 the pay-off she expects at node 1 from going up at that node is lower than her pay-off from 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. As we show in Appendix B below, this proof that it is irrational not to terminated the game at the first node can be extended not merely to the Centipedes of any length, but indeed to all BI-terminating games.[7]

[7] In its general outline, but not in its details, this argument for defection in the Cen-

Thus the Centipede should be terminated in the first move, even though both players would profit if they let the game continue to the third node. This part of the old riddle still applies then. But its other, more important part no longer applies: Our weak form of backward-induction reasoning does not rely on the implausibly strong assumptions that lie behind the standard backward-induction argument.

Note that the proof above only shows that termination is rationally required at the first node. It is not a proof that terminating moves, which are prescribed by the standard backward-induction argument, are rationally required at subsequent nodes. Indeed, given that Alice's move up at node 1 is irrational, our weak assumptions do not imply that Bob at node 2 would have trust in Alice's rationality. And without this trust, we cannot establish that it would be irrational for Bob to go up at node 2. This observation will be relevant to our discussion in the next section.

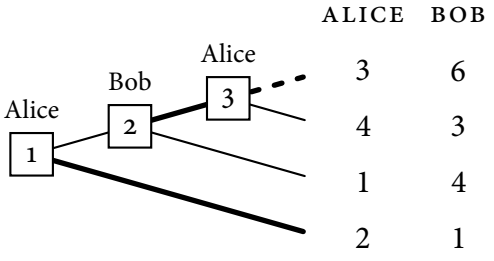
3. A new riddle

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.[8] And assume

tipede is similar to the one in Broome and Rabinowicz 1999. See also Rabinowicz 1998, pp. 108–9 and Aumann 1998, p. 103. (We are indebted to Caspar Hare for his insistence that we clarify all the steps in our proof and, not least, the precise assumptions we need to make about the players' beliefs.) 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 choosing 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 yet undermined.

[8] This is one possible policy for revising beliefs the players might have. But, clearly, other policies also are possible. The players might instead, for example, treat Alice's irrational move at the first node as a momentary lapse on her part and therefore expect that she would act rationally at node 3 and go down. (This is the kind of revision policy that defenders of the standard backward-induction argument might appeal to in

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's irrational to go up 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.



At node 2, Bob 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, 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's enough if the probability of her 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 rational for him to go up at node 2.[9])

---

support of the strong assumption about persistence of trust in future rationality at all choice nodes of the game, even those that can only be reached by irrational play. Compare Sobel 1993, pp. 121–6; 2005, pp. 436–40.) Indeed, in some games that — unlike the Centipede — aren't BI-terminating and that in addition are sufficiently long, there might be room for re-interpreting the seemingly irrational cooperative move on the part of the first player as an invitation to mutually beneficial cooperation. Thus, when considering finitely iterated prisoners' dilemmas, Stalnaker (1998, p. 53) suggests that the second player might well retain his belief in the first player's rationality if he interprets the first player's behaviour along these lines. But this understanding of the situation cannot be upheld in our short Centipede. Since it's irrational for Alice to cooperate (that is, to go up) at node 3, Bob can't expect her to do it and yet retain his belief in her rationality.

[9] But, even with this modification, one might wonder whether it is psychologically

As we have already shown, it is irrational for Alice to go up at node 1. But note that she can predict that, if she were to go up, 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*.[10]

Note that this paradox is not an argument that it's not irrational to go up at node 1. If going up at node 1 weren't irrational, it would no longer yield a higher expected pay-off than going down at that node.[11] Above, we have shown that weak assumptions about rationality and trust in rationality suffice to prove that Alice would be irrational if she were to go up at

---

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 story about how Bob can be expected to react to Alice's irrational move at the initial node, we do not aspire to psychological realism. For our purposes, it is enough if the story is consistent with the weak assumptions about the players' beliefs at nodes reachable without irrational moves that we made in the preceding section.

[10] 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 that underlie 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 pose a significant challenge to our conception of rationality, since an alternative resolution to the single-person version is that the problem arises in the first place because the agent's cyclical preferences (or other non-standard preferences) are irrational.

[11] Our earlier proof established that going up at node 1 is irrational. 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 not irrational for Alice to go up at node 1, then her pay-off would predictably 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 irrational for Alice to go up at node 1, then her pay-off would predictably be higher if she were to go up than if she were to go down.

Claims (i) and (ii) are consistent. Nonetheless, while these results are compatible as they stand, they lead us to a riddle that we now are going to present.

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.[12] But how is it possible? If rational behaviour consists in maximization of expected pay-off, then going down at node 1 is also irrational. But surely it cannot be that every move at Alice's disposal at that node is irrational — that is, rationally prohibited? So, given that we reject the possibility of rational prohibition dilemmas (that is, nodes where all options are rationally prohibited), we seem to have arrived at a contradiction.[13] This is our new riddle of backward induction.[14]

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 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

[12] 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.

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

[14] 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 predictably higher pay-off if she chose to go up at that node.

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. Given these weak assumptions, it no longer follows that Alice would go down at node 3, nor that Bob would expect it and therefore himself go down. It is perfectly compatible with the weak assumptions that Alice would go up at node 3, that Bob at the second node would expect it (because of Alice's irrational move at node 1) and for this reason himself go up. Bob's move up could well be expected by Alice at node 1, thereby making her irrational move up at that node advantageous to her.

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 irrational for Alice to go up at the first node, implicitly relied on the following principle:[15]

> *Dominance* At a node where an agent *S* has finitely many available options, an option 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 option.[16]

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

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

[16] 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. Given Dominance, this argument implies that it would be irrational for Alice to go down. But, if there can be no rational prohibition dilemmas, it cannot be that both moves at Alice disposal are irrational.

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

> *Conditioned Dominance* At a node where an agent 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 y 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 effects, actual and expected, of the option's modified rationality status on the beliefs of the players and thereby also, eventually, on their behaviour at subsequent choice nodes.[17]

The main idea behind this amendment of Dominance is that having a less preferred expected outcome than some alternative option y doesn't necessarily make an option x irrational: it doesn't make it irrational if y's attractive outcome is essentially dependent on y's irrationality. Option y, such as Alice's going up at the first node, on the hypothetical assumption that y is not irrational, may be shown to give the agent a lower expected pay-off than its alternative x (Alice's going down) on the hypothetical

[17] 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 — with its rationality status revised 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.

assumption that x is not irrational.[18] And yet, at the same time, it can be argued that y would give Alice a predictably higher pay-off than x if we in this argument start from the recognition that y is irrational.[19] This allows us to avoid the apparent inconsistency between the earlier proof (in section 2) and the argument that followed (in section 3).[20]

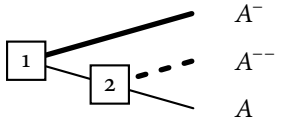
But can we give up the intuitively compelling Dominance? Its hold on us is hard to shatter. It is therefore not obvious that the contradiction we have pointed to can be avoided. We leave this question to the reader.

An alternative solution would be to retain Dominance and avoid inconsistency by allowing rational prohibition dilemmas. Then we would need to accept that all moves that are at Alice's disposal at the first node are rationally prohibited (that is, irrational). We find this harder to swallow than replacing Dominance with its close variant, Conditioned Dominance. Indeed, this solution would incur a further cost. Dominance is part and parcel of a package that also contains the following principle: An op-

[18] Indeed, we have already shown, in the course of our proof in section 2, that Alice's move up at node 1, if assumed not to be irrational, has a lower expected pay-off than Alice's move down at that node. (Note that, since the latter move terminates the game, its expected pay-off doesn't depend on its rationality status.)

[19] As the reader can check, if — as we assume — Alice's irrational move up at node 1 would lead Bob to expect that Alice would also act irrationally at node 3, Conditioned Dominance still suffices to establish that it would be irrational 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 predictably advantageous to her.

[20] Another advantage of Conditioned Dominance as compared with Dominance is that the former — combined with the weak form of backward induction — rules out the following money pump, where A^- is souring of outcome A (that is, it is the same as A except for a small payment) and outcome A^{--} is a souring of A^-:



(This case is structured like the Professor Procrastinate case in Jackson and Pargetter 1986, p. 235, which is predated by a similar case in Bergström 1968, pp. 165–6.) Given Dominance, it is irrational to go down at node 1 if the agent predicts that they will irrationally go up at node 2. By contrast, Conditioned Dominance and the weak form of backward induction imply that going up at node 1 is irrational. More precisely, this holds for a player who (unlike Professor Procrastinate) has no prior record of irrational behaviour. For such a player, who does not expect to act irrationally at node 2, it would be irrational to go up at node 1.

tion is rationally required (permitted) if its expected pay-off is (weakly) preferred to that of each other available option. Given this principle, it would follow that Alice's move up at the first node is rationally required. And yet, as we have shown, this move is also rationally prohibited (irrational). That an option can be both required and prohibited seems even harder to accept than that all options at the agent's disposal are prohibited.[21]

Yet another alternative solution to our riddle would be to reject some of the assumptions that lie behind the weak form of backward-induction reasoning. Maybe our riddle shows that even the weak form of backward induction makes too strong assumptions. If we reject Trust in Rationality, Belief in Trust in Rationality, Introspection, or Logical Competence, then we can no longer prove that it would be irrational for Alice to go up at node 1. This would make the riddle disappear, and revising Dominance wouldn't be needed.[22] But this alternative solution seems rather unattractive. While the assumptions behind the weak form of backward-induction reasoning might not be generally valid, they are plausible enough to be satisfiable in at least some cases, including the case at hand. This possibility suffices for restoring the riddle. Thus, this attempt to make the riddle disappear doesn't seem to be promising.

The solution to the riddle we have suggested — replacing Dominance with Condition Dominance — is in our view more satisfactory than its alternatives. Still, we should be clear about the cost it incurs. This solution implies that, *sometimes, irrationality pays* — that an irrational action might sometimes be predictably more advantageous to an agent than its alternatives. And that's not despite its irrationality, but because of it.[23]

[21] If Dominance is replaced by Conditioned Dominance, then the corresponding changes are needed with regard to the principles concerning rational permission and requirement. We now need to assume that an option x is rationally required (permitted) if, on the assumption that x is not irrational, its expected pay-off is (weakly) preferred to that of each alternative option y, on the assumption that y is not irrational. This principle implies that Alice's going down at the first node is rationally required, just as we would expect it to be.

[22] We are indebted to Tomi Francis and Erik Mohlin for pressing us on this point.

[23] It might be noted that our solution also has implications for decisions under certainty. Suppose we assume, admittedly very unrealistically, that Alice is certain that Bob would go up at the second node if she were to go up at the first node. In this case, not only Dominance, but also Statewise Dominance would imply that it is irrational for Alice to go down at the first node, where Statewise Dominance is the following condition:

Appendices

A. The Irrational Man

The Irrational Man is a classical rationality paradox, due to Haim Gaifman.[24] It was slightly modified by Robert C. Koons, and then additionally modified by Vann McGee.[25] Especially McGee's version exhibits striking similarities to our new riddle of backward induction. It goes like this:

I have a choice between *A*, an empty box, and *B*, a box containing $100. I am promised, by a reliable promisor, that if I choose *A and* this choice is irrational (but not otherwise), I will receive $1000 as a bonus.

It might seem that there is a contradiction here — if we assume that at least one option in the case must not be irrational. My choice of *A* must be irrational, for if it weren't, it wouldn't be rewarded and thus would give me a lower pay-off than if I had chosen *B*. But this would make it irrational. On the other hand, if my choice of *A* is irrational, then (given our assumption above) it must not be irrational to choose *B* instead. But if the choice of *A* is irrational, it would be amply rewarded. Surely, it must

---

> *Statewise Dominance* At a node where an agent has finitely many available options, an option x is irrational if there is some alternative option y such that x's outcome is less preferred by the agent than that of y in every state of nature to which the agent assigns positive credence.

To avoid this undesirable implication, we should replace not only Dominance but also Statewise Dominance by its conditioned version:

> *Conditioned Statewise Dominance* At a node where an agent has finitely many available options, an option x is irrational if, on the hypothetical assumption that x is not irrational, there is some alternative option y such that, on the hypothetical assumption that y is not irrational, x's outcome is less preferred by the agent than that of y in every state of nature to which the agent assigns positive credence.

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

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

be irrational to choose *B* if I would receive more had I chosen *A*?[26]

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 irrational 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 irrational if its predicted outcome were less preferred than that of *A* on the assumption that *A* is not irrational. But, on that assumption, choosing *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 choosing *B* is rational in the case at hand, and that it would be irrational to choose *A*, even though the latter option has a preferred predicted outcome.[27] 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 up, even though the latter move has a preferred predicted outcome.[28]

This analogy, however, should not hinder us from recognizing an im-

[26] 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 must be irrational to accept this bet, because, if it weren't, you would incur a loss by accepting it. And yet, accepting it, while irrational, would give you a more preferred outcome (1 util instead of 0). This seems to make rejecting the bet irrational, but if there are no rational prohibition dilemmas, it cannot be that both your options are irrational. 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 may not be obvious how any analogous 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.

[27] 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 not irrational. Therefore, Conditioned Dominance implies that *A* is irrational.

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

portant difference between the game-theoretic riddle and the Irrational Man. Unlike the former, the latter paradox arises from letting the rational status of options be part of the very specification of its outcome. There is something highly artificial about cases in which it is explicitly stipulated that the option would result in a better outcome if and only if it would be irrational to perform it. One may be tempted to suspect that cases like this contain some internal inconsistency.[29] By contrast, it is decidedly more clear for our game-theoretic riddle, that the underlying decision problem

---

> *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 of the available options is not morally wrong, Consequentialist Dominance leads to a contradiction. 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.)

[29] For example, one might well wonder how the promisor can know whether taking box A is irrational, given that both the assumption that it is and that it is not lead to a contradiction (provided that it can't be that both options at the agent's disposal are irrational). But, if the promisor can't know this, then how can he be relied on when it comes to rewarding irrationality? This problem is avoided when Dominance is replaced by Conditioned Dominance. Then both the agent and the promisor can easily determine that taking box A *is* irrational. Another way to express the worry about the potential inconsistency of the Irrational Man is this: If rationality is equated with maximization of expected pay-off, then the promise to reward taking box A if and only if that action is irrational implies that taking A has a higher expected pay-off than taking B if and only if it has a lower expected utility than that option. This paradox is avoided if rationality instead is equated with the maximization of conditioned expected utility — conditioned on the hypothetical assumption that the option under consideration is not irrational.

is consistent.[30]

B. BI-terminating games

A finite extensive-form game with no moves by nature and with perfect information is *BI-terminating* if and only if standard backward induction prescribes that it be terminated at each of its choice nodes. Another way of characterizing such games is that there is a terminating move m at each choice-node such that if the player S at that node lets the game continue to any immediately succeeding node, and the next player T at that succeeding node chooses a terminating move that gives T a maximal pay-off, then S's pay-off from that move by the next player is lower than S's pay-off from m. Therefore, in a game like this, standard backward-induction recommends terminating the game at each node.

It's well-known that the standard backward-induction argument rests on implausibly strong assumptions. It assumes that the players at all choice nodes of the game would make rational moves and have trust in all players' future rationality. This means it assumes rationality and trust in rationality even at the nodes that cannot be reached by rational play.

Nevertheless, to prove that *it would be irrational not to terminate a BI-terminating game at the initial node*, these implausibly strong assumptions are not needed. It suffices to make relatively weak assumptions that say nothing about how the players would act or what they would believe at nodes than can only be reached if some players were to act irrationally.[31] Here are the assumptions that we are going to rely on for the proof:

> *Dominance* At a node where an agent S has finitely many available options, an option 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 option.[32]

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

[31] Though note that these weak assumptions only imply that it would be irrational not to terminate at the *initial* node. They do not imply that it would also be irrational not to terminate at each subsequent node, as the standard backward induction argument would have it.

[32] Alternatively, we can instead assume Conditioned Dominance. The proof goes through either way.

Trust in Rationality If a choice node c is reachable without anyone making irrational moves, then the player at the immediately preceding node would believe that the player at c would not make an irrational move at c.

Trust in Past Rationality At each node that is reachable without anyone making irrational moves, the player at that node would believe that this node has been reached without anyone acting irrationally.[33]

Introspection At each node that is reachable without anyone making irrational moves, for each move available to the player at that node, if that move is not irrational, then the player at that node would believe it is not irrational.

Logical Competence At each node that is reachable without anyone making irrational moves, the player at that node would believe in whatever logically follows from their beliefs.

Common Belief It is a common belief among the players at nodes that are reachable without irrational moves that Dominance, Trust in Rationality, Trust in Past rationality, Introspection, and Logical Competence hold.[34]

Consider a BI-terminating game in which the longest branch (that is, the longest path through the game tree) consists of n choice nodes. A node c will be said to be of degree i ($1 \leq i \leq n$) if and only if the path that starts with the initial node and ends with c consists of i choice nodes. We are going to prove, by induction, that for any $i > 1$, no node of degree i can be reached without irrational moves.

[33] 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.

[34] In other words, (i) at each node that is reachable without irrational moves, its player would believe at that node that Dominance, Trust in Rationality, Trust in Past rationality, Introspection, and Logical Competence hold; (ii) at each node that is reachable without irrational moves, its player would believe at that node that (i) holds; (iii) at each node that is reachable without irrational moves, its player would believe at that node that (ii) holds; and so on. Note that it follows that a player at a node reachable without irrational moves would believe each clause of this definition of Common Belief and thus, by Logical Competence, would believe in Common Belief.

Base step: Assume, for a proof by contradiction, that a node c of degree n can be reached without any irrational moves. Then, by Trust in Rationality, the player at the preceding node c^- believes that the player at c wouldn't make an irrational move at c. Since c can be reached without any irrational moves, so can c^-. Thus, by Common Belief, the player at c^- believes that Dominance holds. Since it's irrational, by Dominance, to opt for a lower pay-off than one can secure by an alternative move, the player at c^- believes that the player at c would make a move that maximizes that player's pay-off. But, by the characterization of BI-terminating games, that maximizing move of the player at c would give the player at c^- a lower pay-off than they could secure by one of their own terminating moves at c^-. And, by Logical Competence, the player at c^- realizes this. Therefore, by Dominance, it's irrational for the player at c^- to move to c. This contradicts our assumption that c can be reached without irrational moves. Thus we have proved that no node of degree n can be reached without irrational moves.

Inductive step: Suppose that, from our assumptions (Dominance, Trust in Rationality, Trust in Past Rationality, Introspection, Logical Competence, and Common Belief), we have proved that no node of degree $i + 1$ (where $1 < i < n$) can be reached without irrational moves. We now want to prove that the same applies to any node c of degree i. Assume, for a proof by contradiction, that a node c of degree i can be reached without irrational moves. By Trust in Past Rationality, Introspection, and Logical Competence, we have

(i) the player at the immediately preceding node c^- believes that node c can be reached without irrational moves.[35]

Since c^- can be reached without irrational moves, Common Belief implies that the player at c^- believes Trust in Rationality, Trust in Past Rationality, Introspection, Logical Competence, and indeed Common Belief itself. This means that the player at c^- believes the premises of the

[35] *Proof*: If c can be reached without irrational moves, then the same applies to c^-. Hence, by Trust in Past Rationality, (a) the player at c^- believes that this node has been reached without irrational moves. And since c is assumed to be reachable without irrational moves, the move from c^- to c is not irrational. Which implies, by Introspection, that (b) the player at c^- believes that the move from c^- to c is not irrational. Then (a) and (b) imply, by Logical Competence, that the player at c^- believes that c can be reached without irrational moves.

proof that no node of degree $i+1$ can be reached without irrational moves. Hence, by Logical Competence,

(ii) the player at c^- believes that if c can be reached without irrational moves, then moving at c to another choice node is irrational.

Given (i) and (ii), the player at c^- believes, by Logical Competence, that moving at c to another choice node (that is, letting the game continue rather than terminating it) is irrational. And, given our assumption that c can be reached without any irrational moves, Trust in Rationality and common belief in Dominance imply that the player at c^- believes that the player at c will make a move that maximizes that player's pay-off. But, by the characterization of BI-terminating games, that maximizing move of the player at c would give the player at c^- a lower pay-off than one of the terminating moves at c^-. And, by Logical Competence, the player at c^- realizes this. Therefore, by Dominance, it's irrational for the player at c^- to move to c. This contradicts our assumption that c can be reached without irrational moves. Thus we have proved that c cannot be reached without irrational moves.

The above inductive proof establishes that no node of degree i, where $1 < i \leq n$, can be reached without irrational moves. This holds, in particular, for $i = 2$, which means that, at the initial node, it's irrational to let the game continue. This concludes our proof.

Acknowledgements: We wish to thank John Broome, Tomi Francis, Caspar Hare, Harvey Lederman, Erik Mohlin, Robert C. Stalnaker, Robert Sugden, Christian Tarsney, the audience at the philosophy colloquium at MIT on November 3, 2023, and the participants in the workshop on social contract and dynamic choice held at the Institute for Futures Studies in Stockholm, on December 14–15, 2023, for valuable comments.

References

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

Bergström, Lars (1968) 'Alternatives and Utilitarianism', *Theoria* 34:163–170.

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.

Jackson, Frank and Robert Pargetter (1986) 'Oughts, Options, and Actualism', *The Philosophical Review* 95 (2): 233–255.

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.
Sobel, Jordan Howard (1993) 'Backward-Induction Arguments: A Paradox Regained', *Philosophy of Science* 60 (1): 114–133.
— (2005) 'Backward Induction without Tears?', in Daniel Vanderveken, ed., *Logic, Thought and Action*, pp. 433–461, Berlin: Springer.
Stalnaker, Robert (1998) 'Belief Revision in Games: Forward and Backward Induction', *Mathematical Social Sciences* 36 (1): 31–56.
Vallentyne, Peter (1989) 'Two Types of Moral Dilemmas', *Erkenntnis* 30 (3): 301–318.