Idealisation and the Centipede[*]
What is the Significance of the Backward Induction Theorem?

Mats Johansson   Martin Palmé

Department of Philosophy
Lund University
Kungshuset, Lundagård, 222 22 Lund
Sweden
Mats.Johansson@fil.lu.se   Martin.Palme@fil.lu.se

1. Introduction

As we see it the appeal of game theory stems from its ability to supply solutions for a wide variety of decision problems that can be represented as games, in the abstract, theoretical sense of the word. One form of strategic reasoning appreciated for its applicability as well as for its simplicity is the principle of backward induction. This template, however, is not without its problems. In fact, some of its consequences are by quite a few game theorists regarded as counterintuitive and hardly rational at all. We share the opinion that there is a problem. The problem we are interested in has not to do with the logic of the derivation of the backward induction theorem, but with the significance of this result. At the heart of the matter lies the question of what idealised game theory is all about.

In section 1 we introduce the much discussed centipede game, and present some assumptions about the players that are commonly used to establish the backward induction theorem. Section 2 illustrates the principle of backward induction and demonstrates that, on the given assumptions, the game will end at the first node. In Section 3 we describe what we believe to be the gut reaction of a person presented with the backward induction solution to the centipede. Section 4 clears away some extensively discussed issues that we do not regard as relevant to the present investigation. The point of departure for section 5 is a defence of the backward induction solution for the centipede. We readily accept the logic of the argument, but the most important questions remain. In Section 6 we question the significance of the backward induction theorem by noting that no human agent has any reason to act in accordance with the result. This

[*] For helpful comments we would like to thank Linus Broström, Krister Bykvist, Erik Carlsson, Magnus Jiborn, Wlodek Rabinowicz, Nils-Eric Sahlin, Krister Segerberg and Rysiek Sliwinski. We are also grateful for the opportunity to present an earlier version of this paper at a workshop on decision theory and moral philosophy in Uppsala, financed by The Bank of Sweden Tercentenary Foundation.

discussion seems to have important consequences for the enterprise of idealised game theory.

2. The centipede game

A type of game commonly used when analysing the nature of backward induction is the so-called *centipede*. A two player centipede game is characterised by the following conditions: (i) It contains a fixed finite number, N, of decision nodes; (ii) The players alternate making moves; (iii) At a certain node, k, the player has two options: terminating the game with a given outcome – T_k, or continuing the game – C_k (at the last node, both options are in fact terminating moves); (iv) The moving player at node k, P_k, prefers each "remaining" outcome *except T_{k+1} (C_N at the last node)* to T_k.[1,2]

We have chosen the centipede as focus of the discussion, since there is a general agreement that the backward induction theorem, even if it might be contested in other games, at least is valid in this one. The illustration (fig. 1) is a centipede with one hundred basic outcomes.[3]

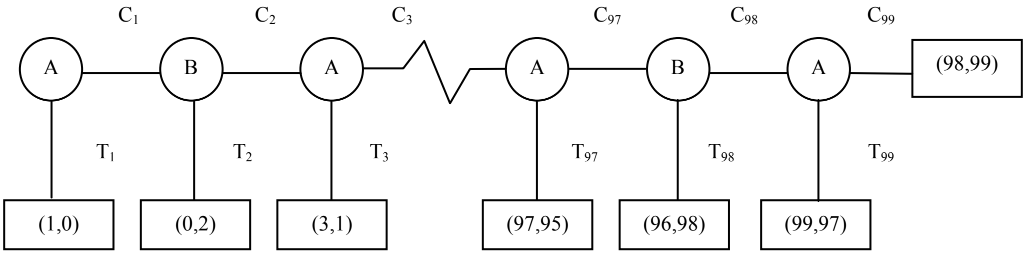

Fig. 1: The Centipede

[1] It is inessential for the backward induction result how P_2 ranks T_1. For the sake of symmetry, we have ranked it as his worst outcome.

[2] Our definition is designed to capture a version of the centipede game. In a general definition, condition (iv) should be: P_k prefers T_k to T_{k+1} (C_N at the last node), and prefers T_{k+2} (if such a node exists) to T_k.

[3] The utility figures are not necessarily on a cardinal scale and there need not be any possibility of interpersonal comparisons.

In order to establish backward induction theorems, a centipede player is supposed to be ‘rational’, which is taken to mean that she (i) chooses as to maximise individual (expected) utility, and thus (ii) has a total, transitive ordering of the outcomes,[4] (iii) has sufficient intellectual resources to be able to reason strategically, etc. The players are also supposed to have certain knowledge: (i’) common knowledge about the game (including knowledge of both players’ orderings), (ii’) common knowledge of both players’ rationality (*CKR)*. This means that the players know each other to be rational, know that both know this, and so on. N.B.: One does not need exactly these assumptions to derive the theorem – other (potentially weaker) principles could suffice. We will return to this issue in sections 5 and 6.

3. Solution by backward induction

Now, what will happen when the centipede game is played? According to the proponent of backward induction (BI), the story roughly goes like this: Player A is the first to make a move, and has the ability to reason strategically. Therefore, she can ask herself: what would happen if the game were at the last node? Well, A would terminate since she is rational and prefers the outcome T_{99} to outcome C_{99}. Player B is, of course, also able to get to this conclusion, and would therefore, if he were at the 98th node, not continue, since he prefers outcome T_{98} to outcome T_{99}. Since A knows that B reasons in this way, she would not want to let him reach the 98th node, … and so on. Going backwards through the decision tree in this way demonstrates that A, if rational, will terminate the game at outcome T_1. [5]

4. “A scandal of game theory”

The fact that outcome T_1 is treated as the rational solution has been called “a scandal of game theory” by Martin Hollis (1998). Is this a reasonable reaction? Let us look at the situation from the “naïve” point of view of a beginner at the centipede game.

When Amy is confronted with the centipede game, she consults her decision analyst friend Calvin who presents her with the BI-argument. Amy, who believes that she and her “opponent” Ben satisfy the rationality and information conditions above, is about to terminate the game at the first node but hesitates. Why do so when the game offers such interesting prospects? It would of course be right to terminate at once if no better outcomes were available. But is it really true that no later outcomes, except T_2, are available? After all, the only risk she

[4] This ordering is supposed to be stable for the duration of the game.

[5] We refer to outcomes via the labels of the actions that bring them about.

takes if she chooses to continue is Ben terminating the game at the next node. Will he do that? It seems as if Ben, just like her, could do better than that. Since there are outcomes further up the decision tree that both players would prefer to the ones immediately available, is it not possible, *rationally*, to reach one of them?

Expressing her thoughts to Calvin, Amy gets the following answer: Your worries can have several sources. But they show that one or more of the conditions above are not fulfilled. Perhaps you doubt that Ben really is a utility maximiser, or you doubt his cognitive resources. Or maybe you doubt whether Ben would trust *your* future rationality, if you chose to continue? It surely can't be denied, that if you accept that you would choose T_{99} if given the chance, and you both know this, etc., the rational thing to do (in fact: the option you *will* choose if the conditions are satisfied) is to terminate the game at the first node.

But Amy is still puzzled. Would it not be strange, she thinks, if one could become a victim of one's rationality and knowledge? If Amy was "less rational", and Ben knew it, he would not have a basis for a BI-argument, and Amy would not have such a basis either. That would make ascension possible! This bothers Amy. How is it possible, she wonders, that both she and Ben could gain by one (or both) of them acting irrationally? Knowing this, Amy asks herself whether or not it is rational to be "rational".[6]

The following is an attempt to explicate and, in a certain sense validate Amy's scepticism about the BI-principle.

5. Changing the game?

It is common in the literature to propose what might well be called *external* solutions to the problem of backward induction. A typical proposal of this kind consists in changing the rules of the game, most often with the motivation that the result is a (socially) more realistic game. Among such "solutions" are: (i) Repeated games, where it may pay to establish an image as a "co-operative" player; (ii) An external force that monitors promises, making it possible to agree on plans without fear of defection; (iii) Games where the number of nodes is not known (or even infinite), removing the base step of the induction argument.[7]

In the present paper, we accept no external solutions – meaning that we will not interfere with the centipede as such. We will, however, to some extent

[6] Binmore (1997) posed, what he called, the "seemingly stupid" question: Is it rational to be "rational"?

[7] Jiborn and Rabinowicz (2000) have shown that, in some cases, it is possible to establish a base step for backward induction without knowing the length of the game.

examine what kind of conditions the idealised agents of game theory should meet.

6. Backward arguments

As we presented Amy's predicament, her most pressing problem at the outset of the game is to rationally evaluate the conditional: "If C_1 then T_2" (since T_2 is the only "threat" if she chooses to continue). And to treat this, she had better come up with some sort of hypothesis about what Ben would do *if* she were to choose C_1. Or so it would seem. Robert J. Aumann (1996) suggests that this line of thought is, in a way, beside the point. Remember Calvin's message to Amy is that if the conditions are fulfilled, she simply *will not* choose C_1.

According to such a position, the whole business of asking how "a beginner at the centipede" would react is misleading and just serves to confuse the theoretical aspects of the discussion. Aumann (1996) attributes the following structure to the BI-sceptic's argument:[8]

> Starting with *CKR* [common knowledge of rationality], we prove that P1 goes down at the first vertex. Now, we say, let's try it again. Must P1 *really* go down at the first vertex? Let's suppose not – i.e., that she goes across. But then we have a contradiction to *CKR*, and anything whatever follows from a contradiction! So we must abandon *CKR*. But then there is no longer any reason to go down at the first vertex. Just as clearly, *this* argument is absurd.

Furthermore, to avoid criticism essentially based on his proof's subjunctive character, Aumann (1998) presents a refined BI-argument with reference to *rationality* only at nodes that are *actually* reached.[9,10]

Now, if the sceptic's argument really should be interpreted as an attempt to refute the inference-relation $CKR \Rightarrow T_1$ (with Aumann's interpretation of *CKR*), the allegation of absurdity might be granted. This, however, is not the case. We believe that very few BI-sceptics want to, in Aumann's own words, "challenge the correctness of his mathematics". Our criticism is, as will become apparent in the next section, aimed at the appropriateness of the representations of rational players and their beliefs proposed by Aumann et al.

Rationality is a contrast phenomenon – choosing one particular option, one thereby chooses *not* to take some other course of action. The basic ingredient in instrumental rationality can be expressed with the phrase "The rational agent

[8] Binmore is the one under attack here.

[9] He proves that "[…] *if at the start of play there is common knowledge of ex post material rationality, then the backward induction outcome results* […]" (Aumann 1998)

[10] See also (Broome and Rabinowicz 1999), whose proof we discuss in section 6.

chooses the option she believes will maximise her personal utility”. If A is rational, we should be able to, from the statement “A chooses T_1 rather than C_1”, infer: “A believes that the option C_1 would (eventually) produce a worse outcome” (provided she is not indifferent between the two options).

It is obviously not impossible to rationalise the behaviour of an agent who acts in accordance with the BI-result. We can, for example, attribute to her the belief that her “opponent” would use an exclusively forward-looking BI-strategy at node 2 (even though the logical basis for such a strategy is questionable). Or she could think that the other player is extremely risk averse, and thus terminates with a small gain as soon as he gets the chance. Or we can stipulate that she for some other reason, based perhaps on a more complex deliberation involving e.g. an estimation of the other player’s *cardinal* utility scale, his attitude towards risk, his aspirations, etc., has come to the conclusion that the probability of T_2 given C_1 is high enough to motivate terminating the game at once.[11]

The fact that it is possible to rationalise BI-behaviour does show that it is indeed rational for the first player to choose T_1 if the *CKR* conditions are satisfied. Still, one might ask what grounds the agent has for her beliefs concerning counterfactual situations. We will not pursue this issue here, since regardless of those grounds, a fundamental question remains: What is the significance of the backward induction theorem?

7. Idealisations

Having studied and accepted the “mathematical” part of Aumann’s argument one might reasonably wonder what it shows us. In this section, we will examine whether or not it is possible to make sense of *BI-theorems* – i.e. theorems that demonstrate that agents equipped with certain sets of beliefs, and satisfying certain rationality constraints will act as if they reasoned by backward induction when confronted with a centipede, and thus will terminate the game at the first node.[12] We conclude that the BI-analysis reveals neither how people actually act (or reason) nor how they *ought* to act when confronted with the centipede.

If the BI-theorem were descriptive ordinary people would, when confronted with the centipede, terminate the game at node one. However, since folk psychology as well as empirical studies indicates that this is not the case, we have no reason to think that ordinary people are the kind of players discussed by

[11] Rabinowicz (1998) discusses rationalisation, with the aim of showing that his rationality conditions does not imply *sustained* common belief of rationality.

[12] We will speak of *the* BI-theorem in the text, though strictly speaking there are of course several such results.

the BI-theorists.[13] Actually, even if people would terminate the game at the first node we would not be in a position to conclude that ordinary people meet the requirements defined in BI-theorems since their behaviour can be explained in a number of different ways. Those who think that the BI-theorem is descriptive in some more subtle way than the one just rejected, must accept the burden of proof.

The theorem could be prescriptive for a person *if* she is the kind of player assumed in the theorem. Terminating the game at the first node would then for her be the only rational thing to do. Although it thus makes sense to speak of a prescription, such a prescription is of little interest to people who seek advice on how to act in the centipede game. If the BI-theorem provides a prescription, it is only relevant to players who meet the conditions set by the BI-theorem. But surely, any interesting notion of prescription in game theory must have implications for ordinary people – people who arguably do not meet such conditions.

Perhaps it makes sense to speak of another kind of prescription, according to which ordinary people ought to become the kind of players assumed in the proofs of the theorem. One could argue that some versions of the BI-theorem idealises the players in a way that makes them good examples of how ordinary people ought to be. There are two problems with this line of thought. First, having reasons to become a certain kind of player does not imply having reasons to act as if one was such a player. Hence, the prescription says nothing of how "less perfect" agents ought to act in the centipede game. Secondly, it is by no means obvious what reasons we have to become the kind of agents assumed in the proof of the theorem. In order to show this, we will examine one crucial assumption made by BI-theorists.

According to the assumption that the players have common belief in rationality, henceforth *CBR*, the players believe each other to be rational, believe that both believe this, and so on. No doubt, *CBR* seems like a rather plausible assumption, at least in the sense that *CBR* in real life is a common assumption, made in order to avoid being exploited.[14,15] The problem, however, is not *CBR* in itself but the additional assumption made by some BI-theorists that the players *sustain CBR* regardless what moves have been made. Hence, even if Amy makes an obviously irrational move Ben remains convinced that Amy is rational. This is clearly a very dubious assumption. Nevertheless, this assumption has been considered essential for the proof of the BI-theorem.

[13] See for instance (McKelvey and Palfrey 1992).

[14] Given of course that there is no information supporting that the other player is irrational.

[15] Another function of assuming CBR might be to avoid wishful thinking.

Quite recently John Broome and Wlodek Rabinowicz (1999) have argued that BI can be established in the centipede with more plausible assumptions than that the players have sustained common belief in rationality throughout the game (*SCBR*).[16] Broome and Rabinowicz suggests that “backwards-induction can be reconstructed for the centipede game on a more secure basis” than *SCBR*. Their alternative is this:

(0) At each round in the game that has been reached without any irrational move, the player at that round acts rationally.

(1) At each round in the game that has been reached without any irrational move, the player at that round believes (0)

(2) At each round in the game that has been reached without any irrational move, the player at that round believes (1).

And so on.

So far, Broome and Rabinowicz have formulated what might seem to be an acceptable idealisation. However, their suggestion has a serious weakness.

In reply to an objection posed by Rysiek Sliwinski, based on the possibility that a player might incorrectly interpret a move as irrational even though it is not,[17] Broome and Rabinowicz instead make some stronger, but purportedly plausible assumptions from which 0, 1, 2, …etc. can be derived. These assumptions are:

(A) At the beginning of the game, both players have no false beliefs

(B) During the game, both players acquire only beliefs that are true

(C) Both players retain all their beliefs so long as they are consistent with their acquired beliefs

(D) At the beginning of the game, there is a common belief in (0), (A), (B) and (C)

[16] Other proofs of the BI theorem, also using “weak” assumptions can be found in (Aumann 1996) and (Rabinowicz 1998) (though Aumann’s proof is in terms of knowledge).

[17] Suppose the game has reached node k with only rational moves, but that P_k wrongly interprets one of the moves as irrational. Then (1) would not be satisfied, which would undermine the BI-argument.

One might reasonably regard Broome's & Rabinowicz' answer as *ad hoc* since they deal with the problem by assuming it away. This however does not bother them, because they regard A, B, C and D to be consistent with traditional idealisations of game theory. What makes such idealisations "ideal" is that they describe the features of players that may be seen as "epistemic 'virtues' rather than vices".[18,19]

Our critique of the appropriateness of A, B, C and D as idealisations, is based on the fact that Broome's and Rabinowicz' answer to Sliwinski makes their account less realistic. Given that people do not meet A, B, C and D, is it possible to argue that they should? One answer, that obviously will not do, is that people gain by meeting A, B, C and D. The reason is this: Since ordinary people do not terminate the centipede until far up the tree, they gain by not meeting the conditions.

It is arguably good to meet A, B and C, but does this imply that we ought to meet D? We think not, since meeting D would most probably damage our motivation to scrutinise new beliefs and reconsider old ones. Hence, we consider D to be something of an epistemic vice. Actually, for agents not meeting A, B or C (i.e. all real life agents), meeting D will also involve having a false belief.

We have seen how the absence of *SCBR* forced Broome and Rabinowicz to accept an equally dubious set of assumptions in order to establish the theorem. As far as we can see, their basis is no more "secure" than the original one. In any case, as we have shown, there is no interesting prescription for a real life agent. She is neither prescribed to act like, nor trying to become, an ideal agent (in their sense).

If the BI-theorem is not prescriptive, is it important in some other way? Let us end this section by briefly examining two attempts to defend this kind of theorising.

One purpose of game theoretical idealisations is to identify, isolate and study interesting aspects of rational deliberation – hoping to learn something by problems of application. In this sense, the BI-theorists could be studying backward induction – a kind of reasoning that actually exists among people. The problem with this answer is obvious. Backward induction is a very simple idea (which is probably one of the reasons why it fascinates so many people). In fact, it seems as if the BI-theorists are not discussing the backward induction principle as such, but an argument that establishes the BI-solution in a highly specific game.

[18] In correspondence with Rabinowicz.

[19] It should be noted, however, that although A, B and D are indeed consistent with rationality, they have nothing to do with the players' being rational.

Another way of defending idealisations made by BI-theorists is to argue that these, like all scientific models, simplify things and give nice mathematical structures to work with. For this we pay the price of less "realism". This defence depends on two points that must be supported: That the model has some explanatory value, and that more realistic models are mathematically intractable. In the present case, the explanatory value seems weak, at the least. Some loss of mathematical beauty can surely be accepted if more important goals of theorising are fulfilled.

8. Concluding remark

We have no quarrels with a person whose level of ambition is nothing but to prove something given something. Still, although the BI-theorist does not directly have to answer to the fact that his theory is neither descriptive nor normative he must answer a fundamental question: What is the point of *discussing* the BI-theorem?

Having argued that the BI-theorem is neither descriptive nor prescriptive in any interesting way, it remains to be shown what it provides us with. It might be claimed that we have an unreasonably restrictive view of the ways in which the theorem might be of value. However, if there is some more subtle kind of description and/or prescription hiding behind the theorem, it remains not only to identify what it is, but also to show that it is adequate.

References

Aumann, R.J.: “Reply to Binmore” *Games and Economic Behavior* 17, 1996

Aumann, R.J.: “Note On the Centipede Game” *Games and economic Behavior* 23, 1998

Binmore, K.: “Rationality and backward induction” *Journal of Economic Methodology* 4:1, 1997

Broome, J. and W. Rabinowicz: ”Backwards induction in the centipede game” *Analysis* 1999

Hollis, M.: *Trust within reason* Cambridge 1998

Jiborn, M. and W. Rabinowicz: “Reconsidering the Foole’s Rejoinder: Backward Induction in Indefinitely Iterated Prisoner’s Dilemmas” in (Rabinowicz 2000)

McKelvey R. and T. Palfrey: “An Experimental Study of the Centipede Game” *Econometrica* vol. 60 1992

Rabinowicz, W.: “Grappling With the Centipede: Defence of Backward Induction for BI-terminating Games” *Economics and Philosophy* vol. 14 1998

Rabinowicz, W. (ed.): *Value and Choice. Some Common Themes in Decision Theory and Moral Philosophy* Lund 2000