The backward induction argument for the finite iterated prisoner's dilemma and the surprise exam paradox

Luc Bovens

There are two curious features about the backward induction argument (BIA) to the effect that repeated non-cooperation is the rational solution to the finite iterated prisoner's dilemma (FIPD). First, however compelling the argument may seem, one remains hesitant either to recommend this solution to players who are about to engage in cooperation or to explain cooperation as a deviation from rational play in real-life FIPD's. Second, there seems to be a similarity between the BIA for the FIPD and the surprise exam paradox (SEP) and one cannot help but wonder whether the former is indeed no more than an instance of the latter. I argue that there is an important difference between the BIA for the FIPD and the SEP, but that a comparison to the SEP can help us understand why the conclusion of the BIA for the FIPD strikes us as a counterintuitive solution to real-life FIPD's.

1. Consider first a loose presentation of the BIA for the FIPD. Players 1 and 2 will play a finite number of runs of a game that has the structure of a prisoner's dilemma. We assume that it is rational not to cooperate in a single-run prisoner's dilemma. But if it is rational not to cooperate in the single-run game, then it is also rational not to cooperate in the last run of the finite iterated game. Whatever your opponent chooses, the non-cooperative choice yields a higher payoff and – just like in the single-run game – there is nothing to be gained over and above the payoff in the last run. Furthermore, if there is no cooperative play to be expected in the last run, then it is also rational not to cooperate in the next to last run, since there is nothing to be gained over and above the payoff in this next to last run. This argument can be repeated back to the first run of the game. Hence, it is rational not to cooperate in any of the runs of the finite iterated game.

This loose presentation of the argument hides a number of crucial steps. I will try to make the argument more rigid. Let a player be *minimally rational*, just in case, if he believes at some run of the game that his opponent will not play cooperatively at any later run, then he will not play cooperatively at that run. The following are the premisses of the argument. There is no cooperative play in the last run of the game and both players believe this at pre-game time. Furthermore, both players are minimally rational and they believe this at pre-game time. If we make certain doxastic

Analysis 57.3, July 1997, pp. 179–86.

assumptions, then it follows that the players will not cooperate throughout the game. First, we assume *closure* – i.e. the players believe the logical consequences of their own beliefs. Second, we assume *retention* – i.e. if a player believes a proposition at some earlier time, then he will continue to believe this proposition at a later time. Third, we assume *transparency* at pre-game time – i.e. if a player believes a proposition at pre-game time, then both players believe that he believes this proposition at pre-game time.

Here is how the argument proceeds. It is given that there will be no cooperation in the last run of the game. In *step 1*, we establish that there will be no cooperation in the penultimate run. Player 1 believes at pre-game time that there will be no cooperation in the last run. By *retention*, he will continue to believe by the penultimate run that there will be no cooperation in the last run. Hence, he will believe that his opponent will not play cooperatively at any later run and, being minimally rational, he will defect in the penultimate run. A parallel argument holds for player 2. It follows that there will be no cooperation in the penultimate run.

In *step 2*, we establish that there will be no cooperation in the next to penultimate run. We need to show first that the players *believe* in the next to penultimate run that there will be no cooperation in the penultimate run. It is a premiss of the argument that both players have a pre-game belief in non-cooperation in the last run. By *transparency*, they have a pre-game belief that both players have a pre-game belief in non-cooperation in the last run. By *retention*, they will continue to believe by the next to penultimate run that both players had a pre-game belief in non-cooperation in the last run. It is a premiss of the argument that the players have a pre-game belief that they are minimally rational. By *retention*, they will continue to believe by the next to penultimate run that they are minimally rational. From the players' pre-game beliefs in non-cooperation in the last run and the players' minimal rationality we derived in step 1 the logical consequence that there will be no cooperation in the penultimate run. Hence, by *closure*, the players also believe by the next to penultimate run that there will be no cooperation in the penultimate run.

The argument can now proceed. Player 1 believes by the next to penultimate run that there will be no cooperation in the penultimate run. He believes at pre-game time that there will be no cooperation in the last run. By *retention*, he will continue to believe by the next to penultimate run that there will be no cooperation in the last run. Hence, he will believe by the next to penultimate run that his opponent will not play cooperatively at any later run and, being minimally rational, he will defect in the next to penultimate run. A parallel argument holds for player 2. It follows that there will be no cooperation in the next to penultimate run.

The argument proceeds in a similar fashion to the conclusion that no cooperation will emerge at any step in the game. Needless to say that this argument has been the source of much discomfort. But is the BIA for the FIPD merely an instance of the surprise exam paradox (SEP)? Do they suffer from the same ailment? To answer this question, let us first turn to the SEP.

2. The SEP is familiar. The following are the premisses of the argument. The exam is to occur on some day of the week (Monday, …, Saturday) at 9 a.m and the student believes this before the week starts – i.e. on Sunday. Furthermore, the exam will be a surprise – i.e. if the exam occurs on some day, then the student will not believe a day ahead of time that the exam will occur – and the student believes this on Sunday. If we make certain doxastic assumptions, then it follows that a surprise exam cannot occur on any day of the week. As before, we assume *closure* and *retention*. In addition, we assume *iteration* or a single-person version of *transparency* – i.e. if the student believes a proposition, then he believes that he believes this proposition – and *memory* – i.e. if the exam has not occurred by some day yet, then the student comes to believes after 9 a.m of this day that the exam has not occurred by this day yet.

Here is how the argument proceeds. In *step 1*, we rule out the last day as a candidate day for the occurrence of the exam. If the exam were to occur on the last day, then it would not have occurred on some earlier day. So, by *memory*, the student would come to believe by the penultimate day that the exam had not occurred yet. By *retention*, the student would continue to believe by the penultimate day that there will be a single exam. Therefore, by *closure*, the student would come to believe by the penultimate day that the exam is to occur on the last day. But then the exam would not be a surprise and it is to be a surprise! Hence, the exam cannot occur on the last day.

In *step 2*, we rule out the penultimate day as a candidate day for the occurrence of the exam. Suppose that the exam were to occur on the penultimate day. We need to show first that the student would then *believe* on the next to penultimate day that the exam cannot occur on the last day. It is a premiss of the argument that the student believes on Sunday that there will be a single exam. By *iteration*, he believes on Sunday that he believes on Sunday that there will be a single exam. By *retention*, he would continue to believe on the next to penultimate day that he believed on Sunday that there will be a single exam. It is a premiss of the argument that the student believes on Sunday that the exam will be a surprise. By *retention*, he would continue to believe on the next to penultimate day that the exam will be a surprise. From the student's belief on Sunday that there will

be a single exam and from the exam being a surprise, we derived in step 1 the logical consequence that the exam cannot occur on the last day. Hence, by *closure*, he also would believe on the next to penultimate day that the exam cannot occur on the last day.

The argument can now proceed. By *memory*, the student would come to believe by the next to penultimate day that the exam had not occurred yet. By *retention*, he would continue to believe by the next to penultimate day that there will be a single exam. We established that he would believe by the next to penultimate day that the exam cannot occur on the last day. Therefore, by *closure*, the student would come to believe by the next to penultimate day that the exam was to occur on the penultimate day. But then the exam would not be a surprise and it is to be a surprise! Hence, the exam cannot occur on the penultimate day.

The argument proceeds in a similar fashion to the conclusion that the exam cannot occur on any day of the week. This result is paradoxical. Suppose the teacher makes an announcement that there will be a surprise exam some time next week and the student believes the teacher. The teacher gives the exam on Tuesday. Certainly the exam would come as a surprise – i.e. the student would not believe on Monday night that there will be an exam the next day. So where did the argument go wrong?

3. I endorse the Wright–Sudbury–Jackson solution to the SEP and will present a concise version here. Recall that, in order to establish that the exam cannot occur on the last day, we had to accept that the student would retain his belief that there will be a single exam until the penultimate day. In order to establish that the exam cannot occur on the penultimate day, we had to accept that the student would retain his belief that there will be a single exam and that the exam will be a surprise until the next to penultimate day. The Wright–Sudbury–Jackson solution distinguishes between two cases: either (i) the student is accurately described as having equally good or better reason to believe that the exam will be a surprise than to believe that there will be a single exam or (ii) the student is accurately described as having better reason to believe that there will be a single exam than that the exam will be a surprise. Suppose that the former description is fitting. If the exam were to occur on the last day, then the student's believing on the penultimate day that there will be a single exam, that the exam will be a surprise and that the exam has not occurred yet entails his believing a contradiction – viz. that he believes and does not believe that there will be an exam on the last day. The student would then have ample reason to abandon the belief(s) which he has least good reason to believe and so he would no longer continue to believe that there will be a single exam by the penultimate day. Hence, the retention of the belief that there

will be a single exam in step 1 of the argument would be unwarranted.

Suppose that the latter description is fitting. If the exam were to occur on the penultimate day, then the student's believing by the next to penultimate day that there will be a single exam, that the exam will be a surprise and that the exam has not occurred yet entails his believing a contradiction – viz. that he believes and does not believe that there will be an exam on the penultimate day. The student would then have ample reason to *abandon* the belief which he has least good reason to believe and so he would no longer continue to believe that the exam will be a surprise by the next to penultimate day. Hence, the retention of the belief that the exam will be a surprise in step 2 of the argument would be unwarranted.

In either case, the student is not justified in projecting his earlier beliefs to the penultimate or the next to penultimate day and, hence, his argument is blocked in either the first or the second step.

4. What about the BIA for the FIPD? Recall that in order to establish that there will be no cooperation in the penultimate run, we had to accept that the players will retain their belief that there will be no cooperation in the last run until the penultimate run. In order to establish that there will be no cooperation in the next to penultimate run, we had to accept that the players will retain their belief that there will be no cooperation in last run and that the players are minimally rational until the next to penultimate run. The belief retentions do not occur within the context of *reductiones ad absurdum*. So the issue is not whether the players are justified in inferring that they would continue to believe these pre-game beliefs to the penultimate or the next to penultimate run *if certain future events were to occur*, but rather, whether the players are justified in inferring that they will continue to believe these beliefs *tout court*.

Under what conditions is one justified in inferring that one will continue to believe some proposition P *tout court*? I propose that (i) one is justified in inferring that one will continue to believe that P, just in case, in all future courses of events which one believes to be *feasible*, one will continue to believe that P; (ii) one believes that a future course of events is feasible, just in case, for all one believes, this future course of events may come about – i.e. it is not the case that one believes that this future course of events will not come about; furthermore, (iii) if one believes that P at some earlier time, one comes to learn some proposition Q in a future course of events such that one has better reason to believe that Q than that P and believing P and Q entails believing a contradiction, then one has ample reason to abandon one's belief that P.

I will now present an argument that *purports* to show that the belief retentions in the BIA for the FIPD are unwarranted for similar reasons as

those offered in the SEP (*cf.* Pettit and Sugden). At pre-game time, the players believe that there will be no cooperation in the last run and that they are minimally rational. Before running the backward induction argument, the players do not believe that a future course of events in which they will have witnessed some cooperative play by their opponents will come about. But, neither do they believe that a future course of events in which they will have witnessed some cooperative play by their opponents will *not* come about. In other words, for all they believe, cooperative play by their opponents may or may not come about. That cooperative play by their opponents may come about is the more feasible, considering that it is in their power to invite their opponents to play cooperatively by initiating cooperative play themselves. Hence, by (ii), the players believe that a course of events in which they will have witnessed some cooperative play by their opponents (COOP) is feasible *and* that a course of events in which they will not have witnessed any cooperative play by their opponents (¬COOP) is feasible.

I take COOP to be the only *candidate* feasible future course of events that could weaken the players' pre-game beliefs. The BIA for the FIPD could then be blocked in the first or the second step. Let us focus on the first step. Suppose that COOP is the future course of events that actually comes about. Then, by the penultimate run, the players reason that their having witnessed cooperative play by their opponents provides them with better reason to believe that their opponents will play cooperatively in the last run than that there will be no cooperation in the last run. By (iii), since their believing by the penultimate run that there will be no cooperation in the last run and that their opponents will play cooperatively in the last run entails their believing a contradiction, they will have ample reason to abandon their belief that there will be no cooperation in the last run. On the feasible future course of events COOP, the players will no longer continue to believe by the penultimate run that there will be no cooperation in the last run. By (i), they are not justified in inferring that they will continue to believe by the penultimate run that there will be no cooperation in the last run. The belief retention in the first step in the BIA for the FIPD is unwarranted and the argument breaks down.

A parallel argument can be constructed for the second step. Suppose that COOP is the future course of events that actually comes about. Then, by the next to penultimate run, the players reason that their having witnessed cooperative play by their opponents provides them with better reason to believe that their opponents will play cooperatively in some later run than that there will be no cooperation in the last run *and* that the players are minimally rational. By (iii), since their believing by the next to penultimate run that there will be no cooperation in the last run, that the players are

minimally rational and that their opponents will play cooperatively in some later run entails their believing a contradiction, they will have ample reason to abandon their belief that there will be no cooperation in the last run *and* that the players are minimally rational. On the feasible future course of events COOP, the players will no longer continue to believe by the next to penultimate run that there will be no cooperation in the last run *and* that the players are minimally rational. By (i), they are not justified in inferring that they will continue to believe by the next to penultimate run that there will be no cooperation by the last run *and* that they are minimally rational. The belief retention in the second step in the BIA for the FIPD is unwarranted and the argument breaks down.

For this counterargument to succeed, certain things must be true about the players' pre-game beliefs. Let us focus on the first step of the BIA for the FIPD. The players believe at pre-game time that COOP is a feasible future course of events. They assign a degree of credence at pre-game time to COOP that is greater than 0, i.e.

(i) $\text{Prob}(\text{COOP}) > 0$.

The crucial move in blocking the argument in the first step is that the players can no longer retain their belief that there will be no cooperation in the last run (NCLR) after having witnessed COOP by the penultimate run. For this to be possible, the following must hold:

(ii) $\text{Prob}(\text{NCLR}) > \text{Prob}(\text{NCLR} \mid \text{COOP})$.

Now consider the following fact of probability theory:

(F) $\text{Prob}(\text{NCLR}) = \text{Prob}(\text{NCLR} \mid \text{COOP})\text{Prob}(\text{COOP}) + \text{Prob}(\text{NCLR} \mid \neg\text{COOP})\text{Prob}(\neg\text{COOP})$.

In order to respect (i), (ii) and (F), Bayesian rational players must assign at pre-game time a degree of credence to NCLR that is smaller than 1, i.e.

(iii) $\text{Prob}(\text{NCLR}) < 1$,

and their degree of credence in NCLR must be such that it will increase after having witnessed persistent non-cooperation by the penultimate run, i.e.

(iv) $\text{Prob}(\text{NCLR}) < \text{Prob}(\text{NCLR} \mid \neg\text{COOP})$.

A parallel argument can be constructed for the second step. The crucial move in blocking the argument in the second step is that the players can no longer retain their belief that there will be no cooperation in the last run *and* that the players are minimally rational (MR) after having witnessed COOP by the next to penultimate run.

To block the argument in the second step, it must be the case that

(v) $\text{Prob}(\text{NCLR} \,\&\, \text{MR}) < 1$

and that

(vi) Prob(NCLR & MR) < Prob(NCLR & MR | ¬COOP).

This shows that appeals to *retention* in the BIA for the FIPD are illicit only if the players do not have *full* belief in some of the premisses of the argument at pre-game time (i.e. they do not assign degree of credence 1 to the premisses at hand) and do not yet have *saturated* belief in some premisses of the argument at pre-game time (i.e. witnessing persistent non-cooperation would still strengthen their belief in the premisses at hand).

In real-life FIPD's the players typically have neither full belief nor saturated belief in the premisses at pre-game time and the Wright–Sudbury–Jackson solution to the SEP can be extended to the BIA for the FIPD. The players consider it to be feasible that cooperation might emerge and if cooperation emerges, then they will no longer retain their pre-game belief that there will be no cooperation in the last run or at least that there will be no cooperation in the last run *and* that the players are minimally rational. Hence, given such credences, the BIA for the FIPD breaks down. However, if the players do have either full belief or saturated belief in the premisses at pre-game time, then they will retain their belief in any future course of events that they consider to be feasible. Hence, given such credences, the BIA for the FIPD holds up.

In conclusion, the BIA for the FIPD differs from the SEP in that the BIA for the FIPD *does* hold up if the players do have full or saturated pre-game beliefs. But, since full or saturated pre-game beliefs are untypical in *real-life* FIPD's, appeals to *retention* in a BIA for a *real-life* FIPD are just as dubious as in the SEP. And this explains why the conclusion of the BIA for the FIPD strikes us as a counterintuitive solution.[1]

CU at Boulder, CB 232
Boulder, CO 80309, USA
bovens@spot.colorado.edu

References

Jackson, F. 1987. *Conditionals*. Oxford: Blackwell.

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

Wright, C. and Sudbury, A. 1977. The paradox of the unexpected examination. *Australasian Journal of Philosophy* 55: 41–58.

[1] The research for this paper was supported by the Council on Research and Creative Work, CU at Boulder and the Edelstein Center for Philosophy of Science, Hebrew University of Jerusalem. I am grateful for comments and suggestions by Erik Anderson, John Fisher, Stephen Kuhn, Stephen Leeds, Iain Martel, John Nelson, Graham Oddie, Philip Pettit, Howard Sobel and an anonymous referee of *Analysis*.