444 Lecture 3.6 - The Backward Induction Paradox

Brian Weatherson

Plan

- To discuss why backward induction isn’t quite as popular with philosophers as with economists.

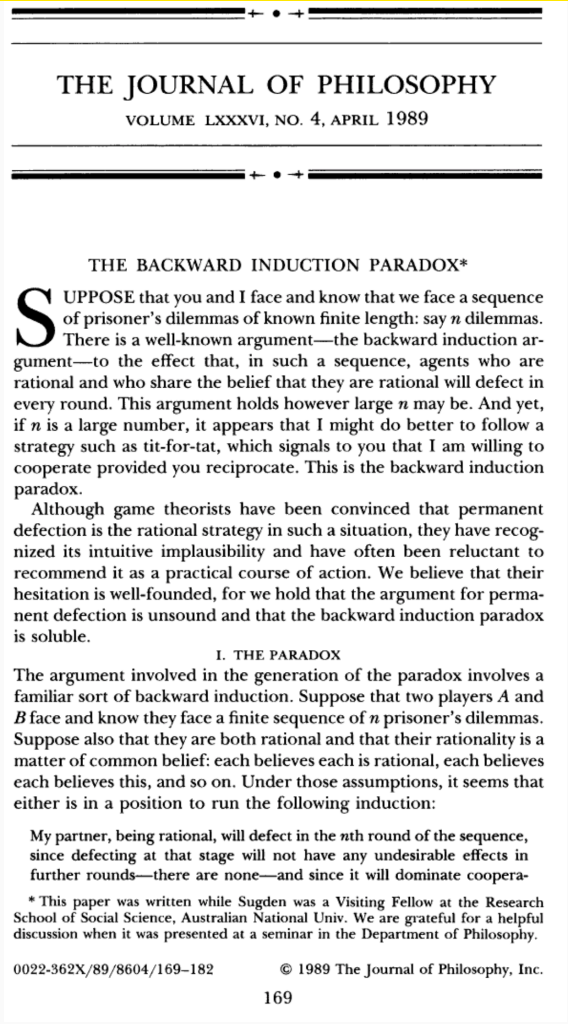
Reading

- No required reading, but if you want to see more, read "The Backward Induction Paradox" by Philip Pettit and Robert Sugden, Journal of Philosophy 1989.

Backward Induction in Economics

- I once heard an economist say the biggest controversy about backward induction reasoning was whether you say "backward induction" or "backwards induction".
- What he meant, and what's true, is that among mainstream economists, this is more controversial than whether the reasoning behind it is sound.
- In philosophy there is somewhat more controversy.

The Backward Induction Paradox



THE JOURNAL OF PHILOSOPHY
VOLUME LXXXVI, NO. 4, APRIL 1989

THE BACKWARD INDUCTION PARADOX*

SUPPOSE that you and I face and know that we face a sequence of prisoner's dilemmas of known finite length: say n dilemmas. There is a well-known argument—the backward induction argument—to the effect that, in such a sequence, agents who are rational and who share the belief that they are rational will defect in every round. This argument holds however large n may be. And yet, if n is a large number, it appears that I might do better to follow a strategy such as tit-for-tat, which signals to you that I am willing to cooperate provided you reciprocate. This is the backward induction paradox.

Although game theorists have been convinced that permanent defection is the rational strategy in such a situation, they have recognized its intuitive implausibility and have often been reluctant to recommend it as a practical course of action. We believe that their hesitation is well-founded, for we hold that the argument for permanent defection is unsound and that the backward induction paradox is soluble.

I. THE PARADOX

The argument involved in the generation of the paradox involves a familiar sort of backward induction. Suppose that two players A and B face and know they face a finite sequence of n prisoner's dilemmas. Suppose also that they are both rational and that their rationality is a matter of common belief: each believes each is rational, each believes each believes this, and so on. Under those assumptions, it seems that either is in a position to run the following induction:

> My partner, being rational, will defect in the nth round of the sequence, since defecting at that stage will not have any undesirable effects in further rounds—there are none—and since it will dominate coopera-

* This paper was written while Sugden was a Visiting Fellow at the Research School of Social Science, Australian National Univ. We are grateful for a helpful discussion when it was presented at a seminar in the Department of Philosophy.

0022-362X/89/8604/169–182

Pettit and Sugden's Paper

That's in part due to this paper.

Iterated Prisoners Dilemma

- It's time to get on the table a game we'll be spending some time on: Iterated Prisoners Dilemma.
- It turns out the central event in the history of the study of this game happened at the University of Michigan, but that's a story for another day.
- A and B will play 100 rounds of the following game.

	Coop	Defect
Coop	3, 3	0, 5
Defect	5, 0	1, 1

Scoring

- This is still a non-competitive game: they are trying to maximise points, not maximise lead over the other.
- But the points add up over all the rounds. (And they don't decay or melt.)
- So each party wants to maximise their sum score over 100 plays of the game.
- At each play, each party knows what the other did on all the previous rounds.
- The strategic form of this is impossibly big; even the two round game has 32 strategies per player, so 1024 cells.

One Shot Reasoning

At any given round, the following reasoning seems sound.

1. If the other player Cooperates, I'm better off Defecting.
2. If the other player Defects, I'm better off Defecting.
3. So either way, I'm better off Defecting.
4. So, I'm better off Defecting.

Repeated Play

But in round one of a repeated game, the following reasoning also looks sound.

1. The best outcome in the long run is if we both Cooperate as much as possible.
2. A plausible way to get that would be to signal that I will Cooperate if, but only if, the other player does.
3. A natural way to implement that is to start Cooperating, then Defect when the other player does (this strategy has become known as Tit-for-Tat).
4. So at round 1 I'll cooperate - if the other player is thinking the same way as me, we'll both make a lot of utility, and relative to how much there is to gain, it's only a small loss if I'm wrong.

Backward Induction

But there is a counter argument.

1. At round 100, there is no signalling value of Cooperating; I just get more from Defecting.
2. Everyone knows this is true.
3. So at round 99, there is no signalling value of Cooperating; the other player will Defect at round 100 whatever I do at 99.
4. Everyone knows this is true.
5. So at round 98, there is no signalling value of Cooperating;...

Temporary Conclusion

- Backward induction suggests that we should defect every round.
- Eventually there will be no signalling benefit to cooperation, and backward induction pushes the moment where that happens back to the start of the game.

Pettit and Sugden

This reasoning is self-defeating.

- Imagine I'm thinking about cooperating for signalling purposes at round one.
- I might worry that the other player will defect come what may at round 2 because of the backward induction argument.
- But the premises of the backward induction argument imply that I'll defect at round 1.
- And at round 2, the other player will know that I did not actually defect at round 1.
- So I should only worry if I think the other player will use an argument whose premises they know to be false.
- And that's not something to worry about.

Short Version

To give up on cooperation requires believing that the other player will think as follows.

- Game theoretic rationality requires defection at every round, so that's what the other player will do from round 3 onwards, so I may as well defect.
- And I know that the other player will do what's game theoretically rational even though they totally did not do that the very last time I interacted with them.
- That's absurd.

Game Theorists Respond

- You should always think the other player is rational.
- If you observe a departure from rationality, you should assume it is a performance error, not a competence error (to use Chomsky's terminology).
- Or, to use the terminology of game theorists, you should assume it was a "trembling hand" error.

A Further Puzzle

- The argument for defecting at round 100 is unaffected by Pettit and Sugden's argument, you should totally defect then.

A Further Puzzle

- The argument for defecting at round 100 is unaffected by Pettit and Sugden's argument, you should totally defect then.
- And I'm not sure that the argument for defecting at round 99 is affected either.

A Further Puzzle

- The argument for defecting at round 100 is unaffected by Pettit and Sugden's argument, you should totally defect then.
- And I'm not sure that the argument for defecting at round 99 is affected either.
- Is round 98 different?

A Further Puzzle

- The argument for defecting at round 100 is unaffected by Pettit and Sugden's argument, you should totally defect then.
- And I'm not sure that the argument for defecting at round 99 is affected either.
- Is round 98 different?
- If you are convinced by their argment that the backward induction argument fails in general, when does it start failing?

For Next Time

We'll end the week looking at a remarkable result involving two player zero sum games.