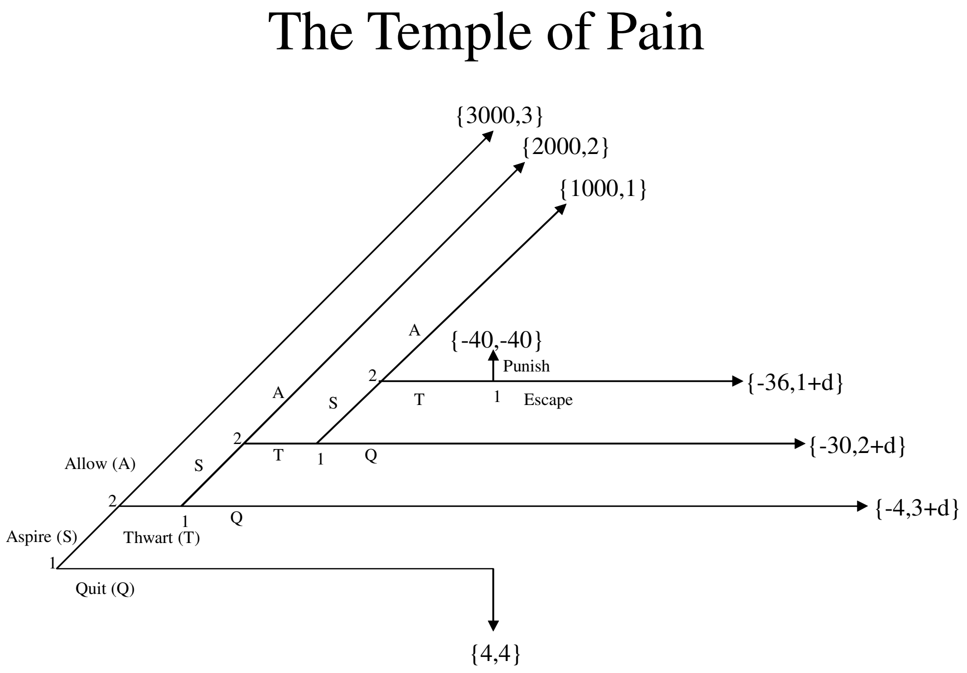
On the Incompatibility of Backward and Forward Induction

Kevin Hasker

14 March, 2016

Consider the following game:



where $1 > d > 0$. At first P1 has the choice of either Aspiring to the large payoff at the top of the pyramid (S) or quitting right away (Q). P2 then chooses either to Allow (A) or Thwart (T) P1's aspiration. This goes on for several rounds (3 in this example) and then P1's final actions are either to Punish P2 by trapping him in the heart of the pyramid or Escaping with P2.

Backward induction picks out only one outcome. P1 chooses Escape at her last decision. P2 then chooses Thwart because $1 + d > 1$—no matter how small d is. Iterating this logic leads to P1 choosing Quit at every other decision node. However, it has other Nash Equilibria. For example the best response to (S, S, X, Y) is Allow (A) because now Thwart results in a payoff of 2 versus a payoff of 3. So one family of Nash equilibria is (Q, X, Y, Z) and (T, α, β), the other family is (S, S, X, Y), (A, α, β). Backward induction picks out one element of the first, Forward induction picks out the second.

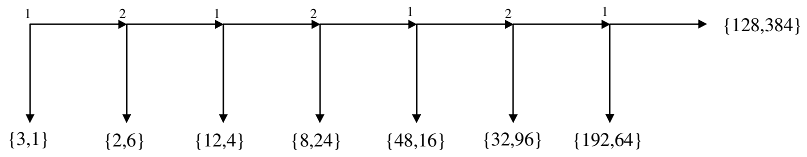
When you see the action Aspire (S), should you not assume that P1 is committed to this action? You know that if she is not committed then it is a mistake, and people often do crazy things for large payoffs, so why not? Because of a trivial d? No, you must be kidding me, if it's small enough what you are doing is stupid. Opening yourself up to "revenge" by P1 for a possibly trivial gain. Why would you do that? You would interpret her actions as commitment, and respond with Allow (A).

While they are not Nash equilibria another benefit of this game is that you can *learn* that P1 is committed. P1 chooses S, and you (being sensible) choose T, now she choose S again, do you choose T? She's shown she's crazy, that the chance of that very large payoff is enough to make her do silly things. So... let's say you are still committed to Backwards Induction, and choose T. Then she chooses S again. Now expected utility might come into your payoffs. She's already given up a certain -4 for a (rational) expectation of -30. At the next decision node (where she's choosing between Punish and Escape) she will only be giving up -36 for -40. You have cost her a chance at a payoff of 3000, are you going to take the risk that she won't choose to punish you? Not me, I would choose A (Allow) and just get it over with. I wouldn't trust her anymore.

So the upshot? Forward induction (loosely defined) can overturn every equilibrium concept you might be fond of. There are competing precise definitions, so for some precise definitions only the Backward Induction equilibria of this game would be allowed, but doesn't $(S, S, S, Punish)$ (A, A, A) seem reasonable?

Let me point out that neither actually works with experiments in the Centipede game:

The Centipede Game



The unlabeled actions are *In* (go across, let the other player make a decision) or *Out*. The experimental results are that people play *In* for a while and then cooperation breaks down. Backward induction says they should always choose *Out* immediately. Forward induction is, of course, vaguer, but does it absolutely contradict rationality? Because only that is needed to choose *Out* at the last decision node. But if it doesn't contradict rationality then iteration leads to the

Backward induction equilibrium. OK, so forward induction respects rationality, but not the iterated version. Unfortunately cooperation rarely lasts to the final decision. So... it doesn't seem that either can explain real behavior in this game.