

This article appeared in Pacific Philosophical Quarterly (September 1999): 278-283)

INFINITE "BACKWARD" INDUCTION ARGUMENTS

Given the military value of surprise and given dwindling supplies and patience, generals are often instructed to (a) attack at a time the enemy does not anticipate, and to (b) attack as soon as possible. These instructions are often common knowledge between attacker and defender. For instance, Saddam Hussein knew the Allied generals were under such instructions at critical junctures of the Gulf war.

Here is an argument (modeled on Sorensen 1993) that did not occur to Saddam Hussein. If the attack could take place during the first hour, then the enemy generals would be obliged to commence the assault during the first hour. But Saddam would then be able to predict the hour of the attack. For Saddam knows that his adversary is obliged to attack as soon as possible. Since the attack would not be a surprise, Saddam can conclude that there will no attack during the first hour. This makes hour 2 the next candidate for an hour in which his adversaries can launch a surprise attack. But the same argument now applies to hour 2. Since Saddam can replicate the reasoning of his adversaries, they cannot find an earliest hour at which a surprise attack is possible. But if there is no earliest hour, then there is no hour at which the surprise attack is possible. (Note the resemblance to Koenig's paradox: there must be a least undefined ordinal because there are more ordinals than definitions. But that very description defines an ordinal!)

The attacking generals can obey their instructions only by scheduling the attack for an hour at which they know Saddam will not

know to be the hour of the attack. But since their instructions are common knowledge, Saddam can know whatever the generals know. In particular, any proof that hour n satisfies the instructions can be replicated by Saddam. One cannot keep a secret from someone who knows everything you know.

Darwinian processes ensure that the disproof of the surprise attack will strike my readers as fallacious. The reasoning may also remind readers of kindred elimination arguments: the surprise test paradox, the iterated prisoner's dilemma, the centipede, etc. These paradoxes are known as "backward inductions". They are mathematical inductions that eliminate the last opportunity in a sequence and then eliminate any opportunity that precedes an eliminated opportunity. For instance, in the surprise test paradox, a clever student argues that the test cannot be given on the last day, Friday. If no test had been given by Thursday, the students would be able to predict the test will be given on Friday because that is the only remaining day. Once Friday is eliminated, Thursday can be eliminated. For if no test is given on Wednesday, the students will know the test is on either Thursday or Friday. By the previous reasoning, the test cannot be on Friday, so they would be able to conclude the test is on Thursday.

Backward induction paradoxes appear essentially finite because they take the last opportunity in a sequence as their base step. However, the sophistry about the earliest surprise attack takes the first opportunity as its base step. It moves forward. This allows Saddam to eliminate infinitely many alternatives:

Base step: Opportunity 1 cannot be the opportunity for the earliest surprise attack.

Induction step: If opportunity n cannot be the opportunity for the earliest surprise attack, then neither can opportunity $n + 1$.

Conclusion: For all n , opportunity n cannot be the opportunity for the earliest surprise attack.

Since the sample space is infinite, our notion of surprise is perturbed by some of the oddities of transfinite probability theory. If there are infinitely many tickets for a lottery with just one prize, then each ticket has a 0 probability of being a winner. Nevertheless, the probability that at least one of them is a winner is 1. Thus, transfinite sample spaces force us to make a syntactic distinction that does not hold for finite sample spaces:

$$(A) \quad (\forall n) \Pr(\text{Ticket } n \text{ is a winner}) = 0.$$

$$(B) \quad \Pr(\exists n)(\text{Ticket } n \text{ is a winner}) = 1.$$

Similarly, if there are infinitely many opportunities for the single surprise attack, then

$$(C) \quad (\forall n) \Pr(\text{Hour } n \text{ is the opportunity for the earliest surprise attack}) = 0.$$

$$(D) \quad \Pr(\exists n)(\text{Hour } n \text{ is the opportunity for the earliest surprise attack}) = 1.$$

The sophistry about the earliest surprise shifts the probability of the existential generalization from 1 to 0.

So one way of obtaining an infinite backward induction is to elaborate on a version that goes forward. A second way appeals to the possibility of an infinite past. Suppose that students at Academy Omega have always been in school. They are scheduled to graduate in the year 2000 AD. Each year, the students are reminded that there will be a surprise test before graduation. (Hence, there have been infinitely many reminders with no first reminder.) After one of the reminders, a clever student objects that the surprise test is impossible:

1. The test cannot be given in the year 2000.
2. If the test cannot be given in year n , then it cannot be given in year $n - 1$.
3. In no year prior to 2000 can the test be given.

This argument goes backward, but since the past is infinite, it eliminates infinitely many alternatives. The argument proves more than would be of practical interest to the clever student. He is propounding the argument at a time that is only finitely distant from the year 2000. For the purposes of prediction, he need only eliminate the finitely many future days that remain as possible test dates. However, the "overkill" should be welcomed by theoreticians.

There are historical challenges to my assumption that an infinite past is possible. The Bible suggests that God created the world finitely long ago (indeed only a few thousand years before Christ). What was God doing before He created the world? "Creating hell for people who ask questions like that" was St. Augustine's most famous answer. But his serious answer was that God created the world immediately. If there were

any time preceding the creation of the world, God would be forced to make an arbitrary choice as to when to commence creation. According to Augustine, to choose without reason is undivine.

Augustine missed a better reason to oppose an infinite past. As a Platonist, Augustine believes that existence is better than non-existence. (This is pivotal in his solution to the problem of evil -- evil is unreal because evil is an absence of being.) Since God is all-good, He prefers to create the world as soon as possible. However, if the past is infinite, God creates the world too late. Therefore, time must have a beginning and God must create the world forthwith. Any time spent deliberating is procrastination. Divine creation must be without forethought. Divine creation precludes divine design.

Religion aside, there is nothing mathematically impossible about an infinite past. We are happy to use the positive numbers to model an infinite future. So, we should be happy to use the negative numbers to model an infinite past. One does not even need an unbounded past. There are infinitely many negative fractions that are closer to negative one than to zero. There could be an infinitely decelerated school schedule that has its last session at 0, its second to last session at $-1/2$, the third at $-3/4$, the fourth at $-7/8$, etc. All the sessions could be crammed into the past minute.

In the finite centipede game, a hundred one dollar bills are placed on a table and two players take turns collecting the money. A player may either take a dollar or may end the game by taking two dollars. Although it seems that the players would be able to collect most of the hundred dollars, there is an argument that only two dollars will be taken. Suppose the game reaches turn 99. Since only two dollars remain, player A has the

choice of taking one dollar or two dollars. Since there is no prospect of further play, he will take two dollars and end the game. Player B would have anticipated this at turn 98. Accordingly, he would have taken two of remaining three dollars, thereby ending the game. But player A would have anticipated this when deciding what to do at turn 97 and so ended it then. The regress continues until we conclude that at turn 1, A would have taken two dollars and ended the game immediately.

An infinite centipede can be constructed by supposing the turns occur on the same schedule as the one minute version of Academy Omega. To avoid the complexities of infinite pay-offs, let us make the pay-off finite with infinitely many parts. In particular, assume that the pay-offs become larger and larger as they approach 0. At turn 0 the pay-off is one million dollars. At turn $-1/n$ the pay-off is $1/n$ of a million dollars. At each turn the player decides whether to take a one dollar bonus for ending the game. If the game were to reach turn 0, player A would take the million plus the one dollar bonus for ending the game (because no further plays are possible anyway). But it is impossible to reach turn 0 because B would have anticipated this at turn $-1/2$. B would have taken his half-million plus the dollar for ending the game because B. For B can make no further profit from the game. Player A would have realized this at play $-1/4$ and so would have taken his quarter million plus the dollar bonus for quitting. But player B whose turn it is at $-1/8$ would have anticipated this. And so on. It is impossible to reach turn $-1/n$ because someone would have ended the game at the prior turn if not before.

There is nothing essentially finite in backward induction arguments -
- and nothing essentially backward. Mathematical induction is essential to

these infinite versions; it is not a mere convenience. If the students were confined to making finitely many applications of inference rules such as modus ponens, then they would have infinitely many disproofs that were collectively exhaustive. But the conclusions and the announcement of a forthcoming surprise test would only form an omega-inconsistent set of statements {`There is an n such that n is the surprise test day', `Day 1 is not the test day', `Day 2 is not the test day', `Day 3 is not the test day',}. The real inconsistency is the simple inconsistency familiar from the finite surprise test paradox.

One might be tempted to further conclude that full common knowledge is needed for infinite backward induction. Finite backward inductions only require that the knowledge operators iterate as many times as there are sub-games in the super-game. But for infinite supergames, the iteration would be infinite and so demand common knowledge in the form deemed unrealistic by many philosophers and economists. Most commentators who reject the game-theoretical backward inductions have singled out the common knowledge assumption as the most suspicious premise of the paradox (for a review of this solution, see Sobel 1993).

However, even for finite backward induction, the last play need not be common knowledge among the players. It is enough if one of the players believe that there is an n such that the other player believes it is common knowledge that the game will stop at n . Even if the belief is false, the backward induction argument seems to show that the player who believes that there is common knowledge will never cooperate. This prediction is paradox enough.

If player A believes that B will never cooperate, then A will also refuse to cooperate. If A is saddled with an uncooperative B, his only recourse is to play defensively. For instance, in an iterated prisoner's dilemma, A must defect at every turn if he believes that B will defect at every turn.

Given this result, we can elevate the dismal result to the next level of iteration: If player B believes that player A believes that player B believes there is common knowledge, then that suffices for B to play defensively in anticipation of A's pre-emptive defensive play. This result sets the stage for the next level of iteration where A anticipates B's pre-emptive pre-emptive defense.

Players A and B might actually agree that there is no common knowledge of the terminating point. They could even agree that common knowledge is objectively impossible. But if A and B lack common knowledge that there is no common knowledge, then their impossibility proofs about common knowledge do them no good. They are doomed when with their partners succumb to the illusion of common knowledge or when their partners fear that their partners have succumbed to the illusion, and so on.

Finite backward induction argument can proceed merely on the appearance of common knowledge. Or even the appearance of the appearance of common knowledge. Or even the appearance of that appearance. And so on ad infinitum. Consequently, the attempt to solve the paradoxes by casting doubt on common knowledge is self-defeating. For the refutation of common knowledge would work only if it were itself common knowledge.

REFERENCES

- Bovens, Luc (1997) "The Backward Induction Argument for the Finite Iterated Prisoner's Dilemma and the Surprise Exam Paradox" Analysis 57: 179-186.
- Carroll, John (1987) "Indefinite terminating points and the iterated prisoner's dilemma" Theory and Decision 22/3: 247-256.
- Sobel, Jordan (1993) "Backward-Induction Arguments: A Paradox Regained" Philosophy of Science 60: 143-133.
- Sorensen, Roy (1993) "The earliest unexpected inspection" Analysis 53/4: 252.