Resolution of the Surprise Examination Paradox

In my view, there are two different set of circumstances that are being confused.

First, I claim that the teacher's statement "You will have a surprise exam next week. The exam is a surprise in the sense that you will not know what day the exam will take place until you come to class and take the exam that same day." is simply false. And it leads to the contradiction in the students' argument. This can be clearly seen in the one day version of the paradox: "You will have a surprise exam tomorrow." is simply a contradiction. Because the teacher's statement is false, the students are mistaken if their argument assumes it to be true and they conclude that there will be no exam. Because, their premises are violated they can only conclude that they do not know whether there will be an exam or not. Just like they did not know before the teacher made the announcement. The teacher giving them a false statement gives them no extra information. Therefore, like Chow [], I claim that the problem lies in the students' interpretation of the teacher's statement.

The problem with the above solution (assuming the teacher's statement to be false) is that, it neglects to address the point that the realization of the exam appears to vindicate the teacher's claim. However, I claim that there is some confusion here. For a statement to be true, it must *always* be true. The teacher's claim is *sometimes* true, but also (necessarily) sometimes false. In fact, the teacher's claim can be made true with probability arbitrarily close to one, but never 1. There is no procedure that the teacher can use that will guarantee he will always be able to "surprise the students" and have the exam that week. If such procedure existed, the teacher would *always* be correct and his statement would be true.

+++++(Consider deleting up to marker)++++++

To make clear that the teacher is always slightly incorrect one should consider the version of the paradox where 5 students stand in line all facing forward, so that the last student can see the back of all others and the first student in line cannot see anyone else. In this version of the paradox, the teacher's claim is that, if he places a stamp on the back of one of the students, the student who has a stamp on his back will not know this until they break formation. Again, the students' argue this can't happen.

If we look at this modified version of the paradox as a game between the students and the teacher, where the teach wins if his statement is correct and looses if it is false. There is a strategy the students can follow that will never let the teacher win. The strategy is as follows:

"Look forward, if you do not see the stamp then it is in your back and you should raise your hand."

This strategy guarantees that the student who has the stamp on his back

always gets it right. All students in front of him will get it wrong. The teacher can claim that the student with the stamp on his back didn't, in fact, know. That he only guessed, just as his peers guessed wrong. But to counter the argument the student who got it right can say that his peers only got it wrong because the statement was false to begin with. In any case, the teacher never wins this game.

In other words, there is no procedure the teacher can use that guarantees that he will always be right. The teacher can be right most of the time and this is what the realization of the exam vindicates. But the teacher *cannot* be right all the time and we cannot therefore conclude that his statement is true.
+++++++++marker++++++++++++

In my view, the paradox is confusing because the *apparent* vindication that the teacher was correct can happen with probability arbitrarily close to one. Therefore, although the teacher's statement is false one should consider the alternative statement:

"You will have an exam next week. And *with probability 99/100* the exam will be a surprise, in the sense that you will not know what day the exam will take place until you come to class and take the exam that same day."

(A procedure for doing this is simple. Randomly assign the exam to the last day with probability 1/100. If the exam is not on the last day, evenly split the remaining probability among the other days.)

The above statement is true, the teacher is not lying. Furthermore, the teacher can make the probability as close to 1 as desired. In fact, the teacher would still be telling the truth if he said that:

"*With probability one*, you will have an exam next week. The exam will be a surprise, in the sense that you will not know what day the exam will take place until you come to class and take the exam that same day."

Notice that we only changed the initial claim by adding the words "with probability one" to the beginning. This statement is also true. If the students had used this interpretation of the teacher's statement, they can rule out Friday (the last day) as a possible exam date also **with probability one**. But that's it, there is no backward induction.

Application to the Prisoner's Dilemma

Now, I believe we are making a similar mistake when applying the same backward induction reasoning to the iterated prisoner's dilemma with a finite and known number of rounds.

From my point of view, part of the confusion stems from the solution concepts used. We need to remind ourselves of what is a *strictly dominant strategy*. Strategic dominance occurs when one strategy is better than another strategy for one player, no matter how that player's opponents may play. This last part is important. A strictly dominant strategy provides the best outcome *no matter what strategy the other player uses*. Therefore, it is *always* best to choose a strictly dominant strategy.

The problem is that when solving the prisoner's dilemma we are making two incompatible assumptions. We assume both that the players are rational and that defection is a dominant strategy. These two assumptions are incompatible and lead to a contradiction. Because of the game's symmetry, if we assume that both players think rationally, we will always find that both players adopt the same strategy, to defect. However, if both players adopt the same strategy then it is better to coperate, a contradiction. There are two solutions to this problem. Either we assume that both players do not think the same way, i.e. one or both of them is not rational, which I do not find very appealing; or we assume that defection is *not* a dominant strategy (i.e. it is not *always* the best thing to do).

I favor the latter option. In my view, defecting is not a dominant strategy in the prisoner's dilemma. It would be if it were always better than cooperating but if your opponent adopts the strategy of "thinking the same way you do" then cooperating is better. Furthermore, we can still assume that defecting is a dominant strategy *with probability one*. This does not lead to a contradiction, so we lose very little. Instead of saying that defecting is a dominant strategy we say that it is a dominant strategy with probability one.

Now let's apply this to the iterated prisoner's dilemma. The backward induction argument for the iterated prisoner's dilemma usually begins with:

"It is always best to defect in the last round."

And then the backward induction ensues. If we change this to "It is best to defect in the last round *with probability one*.". Then we no longer can apply the backward induction and we do not get the weird result that it is always best to defect. We only achieve the conclusion that it is best to defect in the last round with probability one.

A strategy that is suggested by a Nash equilibrium is the best strategy to adopt *if* the other players also adopt the strategies suggested by the Nash equilibrium. If the other players deviate from the Nash equilibrium then this solution concept does not tell us what to do. Considering the Nash equilibrium for the iterated prisoner's dilemma, notice that if one player deviates from the Nash equilibrium by adopting a strategy that has a small but non-zero probability of cooperating in the last round, the Nash equilibrium result does tell us what the other player should do.

In summary, it is NOT *always* best to defect in the prisoner's dilemma. That is false. It is best to defect "almost always".

Similarly for the iterated prisoner's dilemma, it is not *always* best to defect in the last round. It is best to defect in the last round with very high probability, but that's it. We cannot use the backward induction argument because there is an infinitesimal, but non-zero, probability that our opponent will cooperate in the last round.

Application to the Traveller's Dilemma

The analysis is basically the same as above. If you *know* (i.e. it is given) that your opponent is playing the Nash equilibrium strategy of $2 (claiming the minimum value for the broken vase), then you should also play the Nash equilibrium. However, if you do not know what strategy your opponent is using the analysis does not prescribe a strategy to adopt. In particular, if we assume it is *always* best to "never play $100" (the maximum value) then we can use the same assumption over and over to show that it is *always* best to play $2. This is a contradiction because of the word *always*. A counterexample is the situation where your opponent plays, say $95, it is definitely *not* best to play $2 in this case. Because we have a counterexample, it is not *always* best to play $2 and we have a contradiction.

Because of the previous contradiction we conclude that. It is not *always* best to assign zero probability of playing $100. This is analogous to the analysis for the solution of the surprise examination. If your opponent deviates from the Nash equilibrium, it may be in your interest to assign a small but non-zero probability of playing $100. For example, you could play $100 with probability 1/(2^1000) and $99 otherwise. In any case, we *cannot* conclude that it is always best to play $2. We can only conclude that it is best not to play $100 *with very high probability*.

Dimitri DeFigueiredo Jul 18th, 2007.