The Halfers are right in the end: Sleeping Beauty problem

Minseong Kim

Abstract. In this paper, I will examine the representative halfer and thirder solutions to the Sleeping Beauty problem. Then by properly applying the event concept in probability theory and examining similarity of the Sleeping Beauty problem to the Monty Hall problem, it is concluded that the representative thirder solution is wrong and the halfers are right, but that the representative halfer solution also contains a wrong logical conclusion.

1. Introduction: Sleeping Beauty problem

The description of Sleeping Beauty problem appears in Elga (2000), and is given the following:
"Some researchers are going to put you to sleep. During the two days that your sleep will last (Monday, Tuesday), they will briefly wake you up either once (only on Monday) or twice (both Monday and Tuesday), depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will ask you, the sleeping beauty, on what probability you will assign to the outcome of the coin toss turning out to be the Heads. Then they will put you, the sleeping beauty, to back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads?"

2. Introduction: The Thirder - Adam Elga

Elga (2000)'s argument can be summarized as the following:
From now on, let $P(head) = 1/2$ be the unconditional probability that a fair coin toss will produce head. Therefore, $P(tail) = 1/2$.
Given that the result of the coin toss is tail, the probability that the sleeping beauty wakes up on Monday and the probability that the sleeping beauty wakes up on Tuesday should not be different. Therefore,

$$P(Monday|tail) = P(Tuesday|tail)$$

As such (by the definition of conditional probability),

$$P(Monday \cap tail) = P(Tuesday \cap tail)$$

As the result of coin toss only affects what happens on Tuesday (awake, not awake),

$$P(tail|Monday) = P(head|Monday)$$

$$P(Monday \cap tail) = P(Monday \cap head)$$

Thus,

$$P(Monday \cap head) = P(Monday \cap tail) = P(Tuesday \cap tail)$$

As $P(Monday \cap head) + P(Monday \cap tail) + P(Tuesday \cap tail) = 1$,

$$P(Monday \cap head) = \frac{1}{3}$$

Therefore, when first awakened, the sleeping beauty should assign 1/3 to the probability that the outcome of the coin toss is Heads.
The followings are not directly in Elga (2000), but they will aid our discussions:

$$P(Monday) = P(Monday \cap head) + P(Monday \cap tail) = \frac{2}{3}$$

$$P(Tuesday) = \frac{1}{3}$$

Let the unconditional probability that the sleeping beauty is only awakened on Monday be $P(EM)$ (EM represents experiment on Monday only) from now on. $P(EM) = P(head)$ tautologically. $P(ET) = P(tail)$, where ET represents experiment on Tuesday also - meaning that the sleeping beauty is awakened also on Tuesday.
Solving the equations of the following:

$$P(Monday) = P(Monday|EM)P(EM) + P(Monday|ET)P(ET)$$

$$P(Tuesday) = P(Tuesday|EM)P(EM) + P(Tuesday|ET)P(ET)$$

$$P(Monday|EM) + P(Tuesday|EM) = 1$$

$$P(Monday|ET) + P(Tuesday|ET) = 1$$

where $P(Tuesday|EM) = 0$ by the settings of the experiment produces

$$P(Monday|EM) = 1$$

$$P(Tuesday|EM) = 0$$

$$P(Monday|ET) = \frac{1}{3}$$

$$P(Tuesday|ET) = \frac{2}{3}$$

3. Introduction: The Halfer - David Lewis

Lewis (2001)'s argument can be summarized as the following:
Because no new information has been presented to the sleeping beauty, when awakened, the sleeping beauty should assign $1/2$ to the probability that the outcome of the coin toss is Heads. Because $P(head) = 1/2 = P(head \cap Monday)$, $P(tail) = 1/2 = P(Monday \cap tail) + P(Tuesday \cap tail)$ and $P(Monday \cap tail) = 1/4$. Therefore,

$$P(head|Monday) = \frac{1/2}{(1/2 + 1/4)} = \frac{2}{3}$$

$$P(tail|Monday) = \frac{1}{3}$$

The following is not included in Lewis (2001), but deriving them will aid our discussions:

$$P(head|Monday) = \frac{P(head \cap Monday)}{P(Monday)} = \frac{2}{3}$$

$$\frac{1}{2} = \frac{2}{3}P(Monday)$$

$$P(Monday) = \frac{3}{4}$$

$$P(Tuesday) = \frac{1}{4}$$

Also,

$$P(Monday) = P(Monday|EM)P(EM) + P(Monday|ET)P(ET)$$

$$P(Tuesday) = P(Tuesday|EM)P(EM) + P(Tuesday|ET)P(ET)$$

$$P(Monday|EM) + P(Tuesday|EM) = 1$$

$$P(Monday|ET) + P(Tuesday|ET) = 1$$

produces

$$P(Monday|EM) = 1$$

$$P(Tuesday|EM) = 0$$

$$P(Monday|ET) = \frac{1}{2}$$

$$P(Tuesday|ET) = \frac{1}{2}$$

4. Examining both solutions: Events and the Monty Hall problem

It is noticeable that two representative approaches produce different answers for $P(Monday)$ and $P(Tuesday)$. By examining these answers, we may actually see which argument is right. I will argue that both approaches are wrong, but in the end halfers are right. Let me slightly change the question, but this change should not really affect the experiment. Instead of not just waking up the sleeping beauty, the sleeping beauty is told before the experiment is carried out that if the coin toss turns out to be head, the sleeping beauty will be awakened on Monday and be killed. If the coin toss turns out to be tail, then the sleeping beauty will be brought back to sleep with the drug and that be awakened on Tuesday. On Tuesday, the sleeping beauty will just be interviewed and not be killed in any case.
Now let us think about the events. Because the number of events we are considering is considered finite, events can safely be used.
Now $P(tail) = P(survive)$ and $P(head) = P(dead\ on\ Monday)$, unconditionally speaking. But we have two different possible sub-events for the event *survive*: $survive \cap Monday = survived\ on\ Monday$ and $survive \cap Tuesday = survived\ on\ Tuesday$. Should we consider these two events separately? The answer is both yes and no. For Lewis's and Elga's, the answer is yes.
It is noticeable that what Monday and Tuesday actually refer to have not been explicitly defined. There are two possible meaning for $P(Monday)$: $P(Sleeping\ Beauty\ woke\ up\ AND\ today\ is\ Monday)$ and $P(Sleeping\ beauty\ will\ be\ wakened\ up\ on\ Monday)$. In the above, both Lewis's and Elga's arguments use the former definition but without much notice, and this, as we will see, has huge consequences. The latter has probability 1, because the event definitely occurs. Let me first examine the latter definition. Then two sub-events of *survive* are not actually different.

If I survive on Monday (represented by conditional p), then I will survive on Tuesday (represented by conditional q). If I am seen as surviving on Tuesday (q), I should have survived on Monday (p). $p \rightarrow q$ and $q \rightarrow p$, therefore $p \leftrightarrow q$. Invoking Kolmogorov probability theory and representing p and q as sets, p and q have to be the same set. Therefore, these two sub-events are actually tautological to each other. It is wrong to consider them as separate two sub-events that form the event *survive*. The only event that exists is *survive*, not *survived on Monday* and *survived on Tuesday*.

By following this line of definition, it is apparent that the halfer approach is right, the thirder approach is wrong by applying the analogy above. it is shown that these sub-events are not actually separate sub-events, so while $P(Monday|tail) = P(Tuesday|tail)$ is true, this does not mean that $P(Monday \cap head) + P(Monday \cap tail) + P(Tuesday \cap tail) = 1$. What actually happened is double-counting. It should have been just $P(Monday \cap head) + P(tail) = 1$, where $P(tail) = P(Tuesday \cap tail) = P(Monday \cap tail)$, because these three events/sub-events are actually referring to the same event.

We can now see that the case is back to basic coin toss issue, and $P(head) = 1/2$ will be what the sleeping beauty responds when asked for the probability of the coin toss outcome without any new information.

From now on, let us call $P(Monday)$ used in the above case as $P(Wake(Monday))$, or shortly $P(W(M))$, which questions the probability that sleeping beauty would be wakened up on Monday. This is 1 by the construct of the experiment. $P(Wake(Tuesday)$, or $P(W(T))$ is defined similarly.

But what about the former definition of $Monday =$ *Sleeping Beauty woke up AND today is Monday*? After all, this definition is what most have in mind when discussing Sleeping Beauty problem. Are the conclusions that follow from this definition inconsistent with a different definition? I will argue that the conclusions are in fact consistent, that what the thirder argument is committing is basically a variant of Gambler's Paradox.

First of all, let us first calculate what the probability of $P($*Sleeping beauty is wakened up AND today's Monday*$)$, in short $P(SM)$ and $P($*Sleeping beauty is wakened up AND today's Tuesday*$)$, $P(ST)$, assuming that two definitions yield the same conclusion. Going back to the analogy above, $P(W(M)) = 1$ and $P(W(T)) = 1/2$ - in other words, $W(M)$ has two events associated: head and tail, while $W(T)$ has only one event associated: tail. Thus, the ratio between $P(SM)$ and $P(ST)$ is $P(SM) : P(ST) = 2 : 1$. As $P(SM) + P(ST) = 1$, $P(SM) = 2/3$ and

$P(ST) = 1/3$. This conclusion is compatible with the thirder representative approach. So the derivation that assumed the consistency of the halfer approach resulted in seemingly the thirder representative conclusion, and this shows what is wrong in the halfer representative argument.
From this result, thirders may say that $P(SM \cap head) = P(head|SM)P(SM) = (1/2)(2/3) = 1/3$, $P(SM \cap tail) = P(tail|SM)P(SM) = 1/3$, $P(ST \cap tail) = P(ST) = 1/3$, therefore the thirder approach may seem vindicated. We have arrived at the thirder conclusions from the halfer assumption, giving exactly the same set of prediction data derivable from the Elga's paper! But exactly this allows us to see what has gone with the thirder approach, because the thirder conclusion and probability data are successfully derived from the core halfer assumptions. In fact, $P(head|SM)$ is arbitrarily specified as $1/2$, without proper logical reasoning. Sure, intuitively speaking it seems that the information SM does not affect the probabilities of head and tail at all. But the lessons from the famous Monty Hall problem are that this intuitive appeal does not sometimes work.
Let us recall the Monty Hall problem.

> Suppose youre on a game show, and youre given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say No.1, and the host, who knows whats behind the doors, opens another door that is not No.1 which has a goat. He says to you, "Do you want to pick the door that is not your chosen door?" Is it to your advantage to switch your choice of doors? vos Savant (1990)

When the problem is approached with intuition only, it seems that there is no advantage in switching. While No.3 is open and shown as the goat, information about No.1 and No.2 are unknown and therefore it seems plausible to conclude that both doors have equal probability, implying no reason to switch.
But as many know, the answer to this problem is that one should switch.
When the Monty Hall show first begins, each door has $1/3$ probability of having a car. The important point is initial three pair configurations stay on place - car:goat:goat, goat:car:goat, goat:goat:car (No.1:No.2:No.3). For the first configuration, not switching leads to a car, while switching leads to a goat. For the second configuration, not switching leads to a goat, while switching leads to a car. For the third configuration, not switching leads to a goat, while switching leads to a car. Overall, no switching will have a car $1/3$ of the time, while switching leads to a car $2/3$ of the time.
It immediately apparent where $P(head|SM)$ falls into. Let us for now assume

that the coin toss is done on Sunday, right before the experiment. Unconditional probabilities are $P(head) = 1/2$ and $P(tail) = 1/2$. There are three disjoint events that as union has probability of 1: $SM \cap head$, $SM \cap tail$, $ST \cap tail$. These probabilities for now are the probabilities before the experiment. $SM \cap head = 1/2$, as *head* and *tail* are disjoint events and only $SM \cap head$ has any mention of event *head*. Now learning of SM occurred and therefore we may eliminate $ST \cap tail$ for now. Should we now assign $P(SM \cap head|SM) = P(SM \cap tail|SM) = 1/2$? In other words, given the knowledge of SM, should we assign equal probability to *head* and *tail*?
It becomes immediately apparent that this assignment issue is so similar to the Monty Hall problem. We should not. What should be assigned is:

$$P(SM \cap head|SM) = \frac{P(SM \cap head)}{P(SM \cap head) + P(SM \cap tail)} = \frac{1/2}{1 - P(ST \cap tail)}$$

And as long as $P(ST \cap tail) \neq 0$, $P(SM \cap head|SM) > 1/2$.
It is now apparent how the thirder representative approach went astray. It is not supported by proper applications of probability theory that $P(tail|SM) = P(head|SM)$. In fact, it is indeed $P(tail|SM) < P(head|SM)$.
Also notice that Gambler's fallacy cannot be used to support the thirder argument. In Gambler's fallacy, coin toss is done several times. For some first tosses, a gambler notices that more heads appeared for these tosses, so for his next toss, he believes it is more likely that a tail would appear to a head, which is false because assuming fair coin, $P(head) = 1/2$ regardless. But the Gambler's fallacy does not apply to $P(head|SM)$, because the coin toss is not being done several times.
Now let us think about the case where the coin toss is done right after the sleeping beauty is awakened on Monday and asked. In such case, what would happen? The case is simpler here. Because waking up on Monday is completely irrelevant to the coin toss - because the coin toss has not occurred - it will be rational for the sleeping beauty to respond $1/2$ to the probability for the outcome *head* when asked. Otherwise, a future coin toss can only be considered as being entangled to what other irrelevant things have occurred. Here I did appeal to the intuition which I decried above, but if this intuition is not true, then the gambler in the Gambler's fallacy may not be making a wrong decision.
Or we can take a more complex route to arrive at the same conclusion. As said, assume that the coin toss is done right after the sleeping beauty wakes up on Monday. Now, $P(head) = P(head|SM)P(SM) + P(head|ST)P(ST)$. Because $P(head|SM)$ seems to be $1/2$ - because waking up on Monday should not change the probability of the *head* outcome and ST will not oc-

cur if *head* is the outcome, calculation goes: $P(head) = 1/2 = (1/2)P(SM)$. From here, $P(SM) = 1$ is arrived, and this may be cited by the thirders to support their case - because $P(SM) = 1$ is an implausible and impossible value. It is indeed the implausible value and therefore should be rejected. But events may have not been properly defined in this case. Let us check this by invoking Bayes' theorem.

Instead of asking $P(head)$, $P(SM)$ is asked. By Bayes' theorem, $P(SM) = P(SM|head)P(head) + P(SM|tail)P(tail)$ Then we ask: what is $P(SM|head)$? First, note that this probability should equal to the following physical meaning: if the sleeping beauty is awakened and told that the outcome of coin toss is *head*, what should the beauty assign to the probability that SM will occur? The rational answer is 0. Why? Because the coin toss always occurs after SM occurs. The fact that the result of the coin toss is already known implies that SM already occurred before the beauty is awakened - implying that the beauty is awakened for the second time. In other words, $P(SM|head) = 0$ and $P(SM|tail) = 0$. Thus, $P(SM) = 0$ and $P(ST) = 1$. This conclusion follows regardless of what positions one take - thirder or halfer. But of course this conclusion is absurd. There is no way $P(SM) = 0$, because $P(head) = P(head|SM)P(SM) + P(head|ST)P(ST) = P(head|SM)P(SM)$ regardless of the position taken. If $P(SM) = 0$ then $P(head) = 0$, contradicting all positions and whatever is assumed in addition.

Is this, however, really a contradiction in the probability theory? Not at all. In fact, the above can be safely reasoned inside the probability theory. What the above says only is that the event SM is ill-defined in a probability space. But surely the event SM exists!

This suggests that the event SM for the sleeping beauty is in Knightian uncertainty. It is not consistent for the sleeping beauty to include SM into a probability space and only event *head* and *tail* can safely be used, where *head* leads to no ST and *tail* leads to ST. Thus, regardless of whether the sleeping beauty learns that the event just occurred is SM or not, the beauty should answer $1/2$ to $P(head)$ or $P(head|SM)$ - though the latter only works colloquially - just reflecting the fact that the beauty learned fact SM, not formally. When asked for $P(SM)$, the beauty should say probability cannot be assigned and that SM is just uncertain.

One may argue that by creating a variant of $P(head)$ which is $P_2(head)$ that shows the probability of head coming up in the past or future, Knightian uncertainty case might be avoided. But this does not really solve a problem. First of all, Bayes' theorem applies for every "correctly-specified" possible event definable by the experiment. If we accept that original event

head exists and is well-defined, then whatever the case is, SM should be in Knightian uncertainty. Furthermore, what is $P(SM|head-at-past)$, abbreviated as $P(SM|HAP)$? It is of course 0 again. Let us abbreviate $head-at-future$ as HAF, $tail-at-past$ as TAP and $tail-at-future$ as TAF. Thus, $P(SM) = P(SM|HAF)P(HAF) + P(SM|TAF)P(TAF)$. What then probability of SM given head at future coin toss? Of course 1. Because you now know that coin toss in the future! Therefore, $P(SM) = P(HAF) + P(TAF)$. But how do we assign probability to $P(HAF) + P(TAF)$? This in the end returns to the question of assigning $P(SM)$, because the probability of the coin toss occurring in the future equals to $P(SM)$. A circular loop is formed, and therefore this solution does not work.

5. Conclusion

This short paper examined both the halfer and thirder solution for the Sleeping Beauty problem. By analogy and reduction to the Monty Hall problem and by properly applying the event concept in probability theory, it is concluded that both solutions went astray, but in the end halfers are right.

References

1 A. Elga (2000). *Self-Locating Belief and the Sleeping Beauty Problem*, Analysis, **60(2)**, p. 143–147.
2 D. Lewis (2001). *Sleeping Beauty: Reply to Elga*, Analysis, **61(3)**, p. 171–176.
3 M. vos Savant (1990). Ask Marilyn, Parade Magazine:16.