Laying Sleeping Beauty to Rest

Masahiro Yamada
Claremont Graduate University

masahiro.yamada@cgu.edu

1. The Sleeping Beauty Problem

Some researchers are going to put Sleeping Beauty to sleep. During the two days that she is asleep, they will wake her up briefly either once—on Monday—or twice—on Monday and Tuesday—depending on the toss of a fair coin (Heads once; Tails twice). After each waking Sleeping Beauty will be put back to sleep with a drug that erases any memory of the waking.

When Sleeping Beauty is first awakened, what should her degree of belief be that the outcome of the toss is tails? What should her credence be upon learning that it is Monday?

Here is a tempting line of thought. Upon waking Sleeping Beauty is faced with three mutually exclusive and jointly exhaustive possibilities: 1) the coin landed heads and it is Monday, 2) the coin landed tails and it is Monday, 3) the coin landed tails and it is Tuesday. All of these have non-zero probability. When she learns that it is Monday, she can rule out 3) and standard probability theory dictates that the credence for 1) must go up. The only question is how to divide credence among the three options upon waking.

There are two main extant positions, one by Elga (2000) and the other by Lewis (2001) both of which I reject. Here are the credences according to Lewis:

Rational credences according to Lewis

Sleeping Beauty's state	Credence
Upon Waking	$P(\mathit{Heads})=1/2$, $P(\mathit{Tails})=1/2$
Upon Learning it is Monday	$P_+(\mathit{Heads})=2/3$, $P_+(\mathit{Tails})=1/3$

On Sunday, before going to sleep, Sleeping Beauty's credence $P_-(\mathit{Heads})$ that the result will be heads is $1/2$. When she wakes up, she learns nothing new that indicates the coin landed one way rather than the other. So $P(\mathit{Heads})=P(\mathit{Tails})=1/2$. But the updated credences upon learning it is Monday are deeply puzzling. The coin toss could happen before Sleeping Beauty's first waking or after being put to sleep again after the first waking. This will not make any difference to her credences. Suppose the coin toss happens after her being put to sleep again. According to Lewis, Sleeping Beauty should be twice as confident that a fair coin toss in the

future will land heads as that it will land tails. How could this be? It is not as though she receives information from the future through backward causation or something of that sort.

Here is Elga's proposal.

Rational credences according to Elga

Sleeping Beauty's state	Credence
Upon Waking	$P(\textit{Heads})=1/3$, $P(\textit{Tails})=2/3$
Upon Learning it is Monday	$P_+(\textit{Heads})=1/2$, $P_+(\textit{Tails})=1/2$

This avoids the problem of Lewis's proposal. But the initial credences upon waking are deeply puzzling. Sleeping Beauty receives no new information upon waking that she did not have before being put to sleep on Sunday. On Sunday, her credence $P_-(\textit{Heads})$ is $1/2$. How could it be rational to change her credence merely upon waking up?

So on either view, we would have to give up deeply entrenched views about how our credences should and should not change. The majority of those who have weighed in on the matter side with Elga. I shall show that neither Elga nor Lewis is right. There is no need to give up any cherished views to deal with the Sleeping Beauty problem.

2. Sleeping Beauty and fair betting odds

Suppose upon waking Sleeping Beauty is offered a wager. If the result of the coin toss is tails, she wins; otherwise she loses. What are the fair betting odds? Fair odds must be such that there is no net expected gain or loss. I will use the notation *n*:*m* for betting odds that result in a profit rate of n/m in case of a win.

One easy way of seeing what the fair betting odds are is to consider what happens when the bet is repeated multiple times. If the experiment were repeated multiple times in the Sleeping Beauty case, we can expect that two out of three wakings will be in a week in which the result is tails. So the fair betting odds are 1:2; i.e., she makes 50% profit on each occasion she wins. She loses all the money she bets when she loses but since she wins twice as often as she loses, she will come out even in the long term by making 50% profit each time she wins. You might think that these fair betting odds entail that on any given waking she is twice as confident that the result is tails as that it is heads. This would support Elga's view. But this is a mistake. The fair betting odds are indeed 1:2 but this does not mean that Sleeping Beauty's credence is $P(\textit{Tails})=2/3$. To see this, consider the following series of cases all of which are wagers on a roll of dice coming up an even number.

Case 1 When you win (i.e. the roll of dice comes up an even number), you make a profit as given by the betting odds; when you lose, you lose what you put down.

What are the fair betting odds? Since the probability that the result is even is $1/2$, the odds have to be 1:1 (i.e. 100% profit in case of a win). Now consider the following case.

Case 2 When you win, you will be allowed to double your stakes retroactively; e.g., if you initially placed \$1, you will now be allowed to retroactively change that to \$2.

What are fair betting odds? They are 1:2. Suppose you initially put down one dollar. When you win, you are allowed to double the stakes to two dollars. If the odds are 1:2, you will make a profit of one dollar after doubling your stakes. When you lose, your stakes are not doubled so you simply lose the one dollar. Since the probability that the roll of dice will come up an even number is $1/2$, the expected net loss/gain is 0.

We have here a situation familiar from leveraged finance. You can magnify your gains by raising your stakes (often this is done with borrowed money but it need not). What is peculiar about Case 2 is that we have leverage applying only one way: only the gains get magnified. But this too is familiar from finance since it is how a futures option contract works: you get a right (but no obligation) to raise your stakes at a future time. You would exercise such a right if and only if by doing so you can magnify your gains. The lower betting odds compensate for this one-way leverage.

Your credence $P(\mathit{even})$ that the roll of dice came up even is, of course, $1/2$. The fact that you will be allowed to double your stakes if you win is no reason to be more confident of one result of the roll of dice than any other.

Case 3 When you win, a device will be activated that has probability c ($0 \leq c \leq 1$) to result in an audible signal. If there is a signal, you are allowed to double your stakes but not otherwise (e.g. another dice is rolled and a sound is emitted only if it comes up 6. If the sound is emitted you are allowed to double your stakes but not otherwise). You must make up your mind about the betting odds before you place a single bet.

What is your credence $P(\mathit{even})$? It is $1/2$. The fact that there is some probability that you will be allowed to double your stakes is no reason to be more confident of one result of the roll of dice than any other.

What would happen if we also used the device to decide whether you will be offered any bet at all instead of merely to decide whether you will be allowed to double your stakes? That makes no difference, either. If you had to win a lottery to be offered the bet described here, that is no reason at all to be more confident of one result of the roll of dice than any other nor would it change the fair betting odds.

What are the fair betting odds? If c is 0, there is no doubling of the stakes so it would simply be Case 1. If c is 1, you are guaranteed to be allowed to double your stakes if and only if you win so that would be Case 2. We expect the fair odds to go from 1:1 to 1:2 as c grows from 0 to 1. We can be more precise. Let g be the profit-rate in case of a win. Suppose you bet one dollar. When you lose, you lose a dollar and the probability of your losing is $1/2$; so the expected loss is 0.5 dollars. When you win, you will make a profit of g dollars from the first bet and there is some probability, c, that you will make another profit of g dollars. The probability of your winning at all is $1/2$ so your expected gain is $\frac{1}{2}(g + c \cdot g) = 0.5g(1 + c)$ dollars.

The fair betting odds require the expected gain and loss to be equal. Thus, $g = 1/(1+c)$ which means the fair odds are $1:(1+\mathrm{c})$. As expected, if c is 0, the fair betting odds are 1:1, and if c is 1, they are 1:2. The fair betting odds vary between these extremes depending on c. But notice that your credence $P(\mathit{even})$ is fixed at $1/2$ and does not vary with c.

One way stakes can be doubled is if one is allowed to place another bet for the same odds. For instance, a bet might be placed by buying a ticket and you are allowed to buy another ticket. The next is such a case but with a twist. It is the final case:

Case 4 When the roll of dice results in an odd number, the device in Case 3 is triggered once. When the roll of dice results in an even number, the device is triggered twice. Each time the device emits a signal, you are offered a wager on the roll of dice (even you win, odd you lose). However, after each wager your memory of it is erased completely. The probability that any one triggering of the device will result in a signal is c ($0 < c \leq 1$). You must decide on the odds upon being offered a wager.

We have $c > 0$ because you would never be offered a wager if $c = 0$. What are the fair betting odds? They are 1:2. Consider: you might not be offered any bet at all, but if the roll of dice comes up an odd number, you get one shot at an opportunity to place a losing bet; if the roll of dice comes up an even number, you get two shots at opportunities to place winning bets; moreover, each shot has an equal chance of resulting in your placing a bet. If this setup were repeated many times, you expect each losing bet you place to be offset by two winning bets. So the fair betting odds are 1:2 independently of c.

What is your credence $P(\mathit{even})$ upon being offered a wager? Before we do the math, here are some intuitive considerations. If c is less than 1, there is no guarantee that you will be offered a wager at all. So your being offered a wager is evidence that the result of the roll of dice is even since you are more likely to be offered a wager in that case. As c gets smaller, the chances of being offered two wagers diminishes much faster than the chances of being offered a single wager upon the dice's coming up even. So your credence should become more and more like your credence for a case in which the second triggering of the device is canceled if the first one results in an offer of a wager. When you are offered a wager in such a case, there are three mutually exclusive possibilities over which you should remain indifferent: 1) dice came up odd and wager offered, 2) dice came up even and wager offered on first attempt or 3) dice came up even and wager offered on second attempt. So as c approaches 0, $P(\mathit{even})$ should approach $2/3$. On the other hand, if c is 1, you are guaranteed to be offered a wager one way or the other so your being offered a wager is no evidence. Thus, $P(\mathit{even})$ should be $1/2$ if $c = 1$. Now the math. The probability of being offered any wager if the dice comes up odd is c and if the dice comes up even, it is $1-(1-c)^2$. So,

$$P(\mathit{even}) = \frac{1-(1-c)^2}{c+\{1-(1-c)^2\}} = \frac{2-c}{3-c} \qquad (\text{since c} \neq 0)$$

As expected, $P(\mathit{even})$ is $1/2$ if c=1, and approaches $2/3$ as c approaches 0. Unlike the fair betting odds, $P(\mathit{even})$ varies with c. This should not surprise you. Cases 2

and 3 show that credence and fair betting odds can easily diverge depending on the structure of the bet. In this particular case at hand, what matters for betting odds is that you get two shots at placing a winning bet for each shot at placing a losing bet, no matter what c is.

To remove any residual feeling of puzzlement, notice that a rise in $P(\textit{even})$ is coupled with a decrease in the credence that you are offered two wagers given that you are offered a wager. We can see the precise effect of this on the fair betting odds in the following way. Given that the dice came up even, the probability that this is a situation in which you are offered two wagers and hence are able to double your stakes is c^2. So the probability $P(\textit{double})$ that you are offered two wagers given that you are offered one wager is (*not* given that the dice came up even): $P(\textit{double}) = \frac{c^2}{1-(1-c)^2}$. We can now solve the following equation for the profit rate g that is required to have no net expected gain/loss: $1 - P(\textit{even}) = P(\textit{even}) \cdot g(1 + P(\textit{double}))$. Solving this gives us $g = 1/2$, i.e. fair betting odds of 1:2, for all $c \neq 0$. I leave confirming this as an exercise for the reader.

3. The correct credences for Sleeping Beauty

Case 4 is the generalized Sleeping Beauty problem due to White (2006) except that the wager is on the result of a roll of dice rather than a coin toss. If we let c be 1, we get the original sleeping beauty problem. Considerations for Case 4 show that the fair betting odds on a wager that the coin landed tails in the original Sleeping Beauty case are 1:2. This is as it should be: Sleeping Beauty knows that she gets two opportunities of placing a winning bet for each opportunity of placing a losing bet. But these odds do not show that her credence $P(\textit{Tails})$ is $2/3$. Rather, $P(\textit{Tails})=1/2$ as Case 4 shows. This should be obvious. If, as Elga holds, she were twice as confident as not that the coin lands tails and also knows that she is guaranteed to have twice as much at stake when the coin lands tails as when it lands heads, she should be satisfied with betting odds 1:4. Does this mean that Lewis is right and Elga wrong? Not quite.

You start in Case 4. You do not know whether the bet you are placing is your first bet or second. Suppose you are told it is in fact the first bet. This would change your situation to that in Case 3: you are placing your first bet while knowing that there is probability c of being allowed to double your stakes if you win. The new fair betting odds are a function of c and if c is 1 as in the Sleeping Beauty case, they stay the same at $1/2$ when you learn it is your first bet. What about your credence? Your credence $P(\textit{even})$ in Case 3 is $1/2$. The same goes for Sleeping Beauty. But if so, $P_+(\textit{Tails})=1/2$ which is in disagreement with Lewis's proposal. Thus,

The correct credences

Sleeping Beauty's state	Credence
Upon Waking	$P(\textit{Heads})=1/2$, $P(\textit{Tails})=1/2$
Upon Learning it is Monday	$P_+(\textit{Heads})=1/2$, $P_+(\textit{Tails})=1/2$

4. Why this is not incoherent

The above result is intuitively appealing. It has the virtues of both Lewis's and Elga's solutions while avoiding the objections to each. But the credences may seem incoherent. By ruling out it is Tuesday, Sleeping Beauty rules out the last one of 1) Monday and heads, 2) Monday and tails, 3) Tuesday and tails. Once 3) is ruled out, standard probability theory seems to require that the credence for 1) go up; i.e., $P_+(\mathit{Heads}) > P(\mathit{Heads})$.

What is wrong with this thought? We need to notice that the standard credence updating procedure makes an important assumption about the partitioning of possibility space into mutually exclusive and jointly exhaustive partitions. The assumption is usually met as a matter of course, but there are cases in which it is not. To see what the assumption is, consider:

> **Limbo** You are held by a nasty organization in some secret prison system. You know that the prisons are located either in Houston or in Chicago. There is only one prison in Houston (No.1) but Chicago has two (No.2 and No.3). You spend a total of 100 weeks in the prisons. 50 of these are in Houston, 50 in Chicago. When you are in Chicago, you spend the first half of the week in No.2 and the second in No.3. Most of the time you are asleep, but when you are in Chicago they wake you once during the week while you are in No.2, and then another time while you are in No.3. Each time you will be given a drug so you cannot tell whether it is the first or the second time you wake up. When you are in Houston, they wake you once during the week. Moreover, after each week they give you a drug to make you forget which week it is but you do not forget the arrangement itself.

Suppose you wake up. What is your credence that this is a Houston-week? It is $1/2$ since you spend half of the weeks in Houston. Now suppose you are told that you are *not* in No.3. What is your credence that you are in Houston? You might insist that it must go up to $2/3$ because when you wake up you are faced with three mutually exclusive and jointly exhaustive possibilities: 1) you are in Houston, 2) you are in Chicago and in No.2, 3) you are in Chicago and in No.3. The latter two have probability $1/4$ each. Standard probability theory dictates that ruling out 3) results in change of credence for 1) to $2/3$. But this would clearly be silly. After all, which 25 weeks that you spend in Chicago have you ruled out? Obviously none. Each week in Chicago is such that you first spend some time in No.2 and the rest in No.3. The fact that you are not in No.3 tells you that if you are in Chicago it is the first half of the week; but none of the 50 weeks in Chicago have been ruled out.

So what is wrong with the silly reasoning? When we are faced with the task of dividing our credence among several possible candidates of type F, e.g. weeks in Chicago vs. weeks in Houston, we need to partition possibility space in such a way that each partition corresponds to a distinct type of F. Otherwise, ruling out that we occupy a certain region of possibility space will not guarantee that we have whittled down the available options. Normally, this is no problem. If one of the partitions corresponds to a distinct type of F, we usually can expect that a way of dividing up the rest of possibility space that results in jointly exhaustive and mutually exclusive

partitions satisfies this constraint. But what we are seeing here is that in fact not any old way of partitioning into mutually exclusive and jointly exhaustive partitions works: possibilities 2) and 3) are not types of weeks but just fragments of a single type of week. Putting it metaphorically, one might think of the partitions as regions in a three dimensional space and the types of weeks they correspond to as their projections onto a two dimensional plane. Such projections can coincide.

Back to the Sleeping Beauty. Her situation is like yours in Limbo. It is not as though there are three types of weeks: 1) heads and she is woken only on Monday, 2) tails and she is woken only on Monday, 3) tails and she is woken only on Tuesday. Rather, there are only two types of weeks: one in which the result is heads, another in which the result is tails. The latter type of weeks is such that she can be located at either of two mutually exclusive temporal positions when she wakes up. But these temporal positions do not correspond to distinct options she is interested in: both correspond to the coin's landing tails. So when Sleeping Beauty rules out the last one out of 1) Monday and heads, 2) Monday and tails and 3) Tuesday and tails, this does not amount to her ruling out one of the options she is interested in. Sleeping Beauty's credence that the coin landed heads does not change upon learning that it is Monday.

References

Elga, A. 2000. Self-locating belief and the sleeping beauty problem. *Analysis* 60 (2): 143.

Lewis, David. 2001. Sleeping beauty: Reply to Elga. *Analysis* 61 (3): 171–176.

White, R. 2006. The generalized sleeping beauty problem: a challenge for thirders. *Analysis* 66 (2): 114.