Sleeping beauty

PHIL 20229
Jeff Speaks

April 10, 2008

1 The set-up

Sleeping Beauty is told the following:

> You are going to sleep for three days, during which time you will be woken up either once or twice. On Day 1, a fair coin is tossed. If that coin comes up heads she will be woken only on Day 1, if tails then on Day 1 and on Day 2. If she is woken on Day 1, then she will be given a drug to put her back to sleep which also causes her to forget that awakening.

Now suppose that you are sleeping beauty, and you are woken up from your sleep. You know the above, and you know that you are being awoken on Day 1 or on Day 2. What should you think is the chance that the coin flipped on Day 1 came up heads?

The argument for $\frac{1}{2}$ seems straightforward: Sleeping Beauty knows that the coin is fair, and so also knows that there is a $\frac{1}{2}$ chance that it comes up heads on any given throw, and a $\frac{1}{2}$ chance that it comes up tails. She has learned nothing which makes her doubt these probabilities for the Day 1 coin toss; so she should still estimate that there's a $\frac{1}{2}$ chance that the coin came up heads.

This involves some general principle of the following sort:

> If you estimate that the probability of some particular event occurring are n, then, if you learn nothing new relevant to the determination of the odds of that event occurring, you should stick with your estimate that the probability of its occurrence is n.

This principle — which sums up the idea that you should only change your view about the probabilities of events in response to new information about the probabilities of those events — seems almost too obvious to be worth stating.

2 Two arguments for $\frac{1}{3}$

However, there are at least two powerful arguments for the conclusion that Sleeping Beauty is a counterexample to this sort of principle.

Let's let 'T1' abbreviate the proposition that the coin came up tails and it is Day 1, and 'H1' abbreviate the corresponding proposition about the coin coming up heads and it being Day 1. Then the first argument can be stated as follows:

1. $(P(T1\|T1 \text{ or } T2)=(P(T2\|T1 \text{ or } T2)$	premise
2. $P(T1)=P(T2)$	(1)
3. $(P(H1\|H1 \text{ or } T1)=\frac{1}{2}$	premise
4. $P(H1)=P(T1)$	(3)
5. $P(T1)=P(T2)=P(H1)$	(2,4)
C. $P(H1)=\frac{1}{3}$	(5)

This argument requires two assumptions, which are stated in premises (1) and (3). How would you argue for these assumptions, if you were trying to defend the argument? Can you argue for (3) by changing the case so that the coin flip occurs *after* the Day 1 awakening? (See the Elga paper on the web site.)

The second argument runs as follows: first, we can imagine changing the case so that whether or not the coin comes up heads, Sleeping Beauty is awakened both days. So now when Sleeping Beauty is awakened, her probability assignments should clearly be as follows:

$$P(T1)=P(T2)=P(H1)=P(H2)=\frac{1}{4}$$

since there's no reason to favor any of the possibilities over the others. But now suppose that we change the case slightly, so that if it is Day 2 and the coin toss was heads, soon after awakening you are told this fact. Suppose now that you are awoken, and that you are not told this. So you can rule out H2 as a possibility. What probability should you assign to the other possibilities? We'll, it seems that you have learned only that one of four equiprobable theses is false, so you should maintain the view that

$$P(T1)=P(T2)=P(H1)$$

But then, as in the first argument, we can infer that $P(H1)=\frac{1}{3}$.

There's also a kind of intuitive argument for the conclusion that $P(H1)=\frac{1}{3}$. Sleeping Beauty would be reasonable to believe that, were this experiment performed over and over again, she would have twice as many tails-awakenings as heads-awakenenings. So, given a random awakening, she should think that it is twice as likely that it be a tails-awakening as that it is a tails-awakening. So, she should think that the odds of heads having been thrown on any particular awakening of this sort is $\frac{1}{3}$. (Imagine us forcing Sleeping Beauty to bet on whether the coin came up heads on each awakening over a series of trials of the case. Wouldn't she stand to do much better if she adopted as a hypothesis to guide her betting that $P(H1)=\frac{1}{3}$?)

3 The generalized sleeping beauty

In the article on the web site, White considers a different version of the problem, which seems to count in favor of the view that we should say that $P(H1)=\frac{1}{2}$.

Suppose we change the original scenario so that each time you would be awoken in that scenario, you have a $\frac{1}{100}$ chance of being awoken in the new version. So in this new version, when you are awoken, you do acquire genuinely new information: you learn that you were awoken at least once. In this case, how should you estimate the chances of heads versus tails?

You should consider first the probability that you will be awoken, given that heads came up: that is clearly $\frac{1}{100}$.

Now consider the probability that you will be awoken at least once, given that tails came up: in that case, you get two chances at being woken up, so the probability is higher: $1-(\frac{99}{100})^2$.

So what probability should you assign to the proposition that heads came up? Presumably it should be the first of these probabilities divided by their sum, i.e.

$$\frac{\frac{1}{100}}{\frac{1}{100}+1-(\frac{99}{100})^2} = \frac{.01}{.0299} \approx .334$$

So in this case, you should think that $P(H1) \approx .334$.

Here's the interesting part. If we change the chances that you would be woken up each time – say, from $\frac{1}{100}$ to $\frac{99}{100}$ — this result is preserved, but it looks like the degree to which you should prefer the hypothesis that the coin comes up tails changes. In that case, the relevant calculations look like this:

$$\frac{\frac{99}{100}}{\frac{99}{100}+1-(\frac{1}{100})^2} = \frac{.99}{1.9899} \approx .498$$

As the chance of being awoken each time gets closer to 1, the probability that we should assign to the H1 gets closer are closer to $\frac{1}{2}$.

The problem is that, as White points out, the arguments for saying that $P(HJ1) = \frac{1}{3}$ seem to work just the same for every chance. (The original case is just the one in which you have a probability of 1 of being awoken each time.) But this can't be right; so something is wrong with these arguments, White concludes.

4 A parallel to the two-envelope paradox?

Remember that in our discussion of the two-envelope paradox we arrived at the following unsatisfactory-seeming conclusion:

> We are not rational to exchange before any envelopes are opened. But no matter what amount is in my envelope, were I shown it, I would be rational to exchange. So I know that there is some piece of information such that, we I to learn it, it would be rational for me to exchange envelopes. Nonetheless, it is not rational for me to exchange now — even though I know this.

This seems unsatisfactory because it is hard to believe that I could know *now* that there's some piece of information such that, were I to acquire it, it would be rational to do such and such, without it being rational for me to do such-and-such now, on the basis of this knowledge.

But if we think that the right answer to the Sleeping Beauty paradox is that $P(H1) = \frac{1}{3}$, then that also seems to be a case of this sort. For if I am asked before waking up what the probability of heads is, I should clearly say $\frac{1}{2}$ — even if I know all the facts about when I will be awoken. So, in that case, if I am asked what the chances are that I will be awoken on Day 1, I should also clearly say $\frac{1}{2}$.

Now, if the arguments for $P(H1) = \frac{1}{3}$ after waking are true, we know that we will wake up, and we know that when we wake up, no matter what happens it will be rational to believe that $P(H1) = \frac{1}{3}$ — even though it is not now rational to adjust down the probability that the coin will come up heads, and I will be awoken on Day 1.

The analogy is then:

> *Initial view*: expected utility of switching/probabilty that the coin came up heads, and that I will be awoken only on Day 1
>
> *Relevant unknown*: amount of money in envelope/I am awake at *this* time.
>
> *Effect of learning*: expected utility goes up/probability the coin came up heads and that I will awoken only on Day 1 goes down

Could this really be the right thing to say about these cases?