Sleeping Beauty

Sleeping Beauty, as described on wikipedia, citing Arnold Zuboff

Sleeping Beauty volunteers to undergo the following experiment. On Sunday she is given a drug that sends her to sleep. A fair coin is then tossed just once in the course of the experiment to determine which experimental procedure is undertaken. If the coin comes up heads, Beauty is awakened and interviewed on Monday, and then the experiment ends. If the coin comes up tails, she is awakened and interviewed on Monday, given a second dose of the sleeping drug, and awakened and interviewed again on Tuesday. The experiment then ends on Tuesday, without flipping the coin again. The sleeping drug induces a mild amnesia, so that she cannot remember any previous awakenings during the course of the experiment (if any). During the experiment, she has no access to anything that would give a clue as to the day of the week. However, she knows all the details of the experiment. Each interview consists of one question, "What is your credence now for the proposition that our coin landed heads?"

One Self: The Logic of Experience by Arnold Zuboff,
http://www.informaworld.com/index/902035430.pdf

Ways to argue different cases:

The cancellation bet (halfer argument)

Suppose, on Sunday, Sleeping Beauty decides to accept the following bet:

- If the coin lands heads she gets $3
- If the coin lands tails she must pay $2

It's clear that this bet is in her favour, as the coin is fair. However, she will also be given the opportunity to cancel this bet any time she wakes up. If Sleeping Beauty must decide all actions before waking up (e.g. she knows that whatever reasoning she uses now she will use later), clear that she should choose not to cancel the bet. If she cancels she gets $0, if she goes ahead she gets an expected $0.5. It turns out that if SB can choose at random she will cancel $\frac{1}{4}$ of the time (as she gets two chances to cancel when the coin lands tails, which is when she would lose anyway).

Difficuly: Note that sleeping beauty's options are a little more subtle than "when she is woken up she is given the option to cancel the bet" here. This is because she cannot cancel the bet twice. So when the coin lands tails one of:

1. SB cancels bets at random - but now she cancels more often when it's tails, so of course she's happy to bet on heads
2. SB is only given the option to cancel the bet while it's still active. So if she cancels on Monday she can't on Tuesday. On waking, she can then reason "given that I've been offered the chance to cancel this bet, it must be Monday", changing the odds
3. SB is vacuously given the option to cancel her bet even when it isn't on - in this case she can reason, on waking "I may not be canceling a bet if it's

tails, I definitely am if its heads - this makes me cancel more often when I'm winning", pushing her away from canceling, it's biased against her.

Bet initiation (thirder argument)

On the other hand, suppose that SB gets the chance to initiate new bets when she's woken up, keeping the same payouts from before. Now, SB did not want to cancel her first bet. So will she want to take on even more of these each time she is woken up? Actually, no. That is because SB will be woken twice when the coin is tails, and only once when the coin is heads. With the payouts above that works against SB.

Expectations and Probabilities - contradiction?

One might at this point see an apparent contradiction. If Sleeping Beauty doesn't want to initiate a bet on waking up, then one could argue that her expectation for this bet is negative, and write something like:

$$\begin{aligned} 0 > E &= 3P(\text{Heads}) - 2P(\text{Tails} \\ E &= 3P(\text{Heads}) - 2(1 - P(\text{Heads}) \\ E &= 5P(\text{Heads}) - 2 \end{aligned}$$

hence $P(\text{Heads}) < 2/5 < 1/2$.

However, if SB doesn't feel like canceling her earlier bet then we get the reverse: $P(\text{Heads}) > 2/5 > 1/3$! So, we see that at least one of:

1. probabilities may not have intrinsic meaning here, we have a choice of models that look different
2. the conversion from (this is a good bet) $\rightarrow$ expectation is flawed
3. the conversion from expectation $\rightarrow$ probability is flawed

Purported Proof that 'halfers' are wrong, no betting required

Note that we can easily calculate the following quantities:

1. $P(\text{Heads} \mid \text{Monday}) = \frac{1}{2}$ (it's like no fancy experiment is happening, this is what you'd think if you knew it was Monday)
2. $P(\text{Heads} \mid \text{Tuesday}) = 0$

Using these, we can calculate what we're interested in:

$$\begin{aligned} P(\text{Heads}) &= P(\text{Heads} \mid \text{Monday})P(\text{Monday}) + P(\text{Heads} \mid \text{Tuesday})P(\text{Tuesday})^{*} \\ &= P(\text{Heads} \mid \text{Monday})P(\text{Monday}) + 0^{**} \\ &= \frac{1}{2}P(\text{Monday})^{***} \end{aligned}$$

* this is called the partition rule
** because $P(\text{Heads} \mid \text{Tuesday}) = 0$
*** because $P(\text{Heads} \mid \text{Monday}) = \frac{1}{2}$

If you're a 'halfer', then you believe that $P(\text{Heads}) = \frac{1}{2}$, but by the above $P(\text{Monday}) = 2P(\text{Heads}) = 1$.

That's to say, if you're a 'halfer' then you're certain that it's Monday, but the whole point of the experiment is that you're clearly *not* certain that it's Monday. The memory erasing drug means that sometimes when you wake up, it will indeed be Tuesday!

What day is it? Betting on Monday vs Tuesday

In response to the above one might complain that 'what day it is' is the problem, and isn't well defined. To make notions of "what day is it" more concrete we can introduce betting. Suppose SB is asked whether she would like to take up the following bet each time she is woken:

- If the the day is Monday she gets \$1
- If the the day is Tuesday she must pay \$$n$

should she take it? Well, SB will definitely wake on Monday, half the time she will wake on Tuesday. So SBs expectation is: $1 - \frac{n}{2}$. SB will expect to profit from this bet whenever $n < 2$.

One might decide to use this observation, about the bet that SB is considering, to *derive* SBs 'credence' that it is Monday. Though we have seen at the top that this allows us to derive multiple answers, for different betting set ups. Suppose we follow this through in this set-up, We assume that 'A bet is good whenever I have a positive expectation' and 'my expectation is a simple function of the probability of some event'. We can then set $n = 2$, and the calculation would look like:

$$\begin{aligned} E = 0 &= 1P(\text{Monday}) \times -nP(\text{Tuesday}) \\ 0 &= P(\text{Monday}) - 2(1 - P(\text{Monday})) \\ 0 &= 3P(\text{Monday}) - 2 \end{aligned}$$

Hence we would conclude that $P(\text{Monday}) = \frac{2}{3}$.

However, under different betting structures we will get different outcomes.

Probability Models

One view is that: we cannot calculate any of the probabilties of interest without a model. All of the reasoning above implicitly assumes some kind of formal probability model that we are working within. Of course, we could choose any probability model. It's then up to us to check that the way we map from this probability model to the real world corresponds to what we want.

For example: if we say that $P(\text{Heads} = \frac{1}{2}$, that could be because "Heads" in this case represents the event that the coin landed heads from the perspective of the experimenter. if we choose a model where: $P(\text{Heads} = \frac{1}{3}$, this could because we (1) always make the same decisions on waking, and (2) we must make calculations about whether to accept bets proposed to us when we wake (and that will always be propsed). Thus our probability model allows us to account for the fact that each bet has double-weight when the coin lands tails.

It is important, when deciding on a model, that a clear relationship between events in the model and events in the situation under examination is made. Then it's also important to check that the model, and it's interpretation don't give rise to answers that are clearly incorrect.

We usually want models to satisfy certain criteria. Often, as long as they satisfy some specified set of criteria we can see that they will be enough for our purposes. In simple cases, such as modeling a fair coin (and no funny business) there is usually a fixed set of criteria that we always demand that fully specify the behaviour of the coin.

In the case of Sleeping Beauty we found it helpful to certain bets.

When constructing these models, we may want to construct different models depending on which betting scenario we're considering. In the case of the cancellation bet a simple $P(\text{Heads}) = \frac{1}{2}$ model works well. In the case of initiation of bets, reasoning about this may be easier if we choose the probability model which sets $P(\text{Heads}) = \frac{1}{3}$. This is because we've used deductive reasoning, outside the probability model, to make sure that the model we've chosen reflects the outcomes that we're interested in.

What Possible Models for Sleeping Beauty?

There are various properties that we might want our probability model to satisfy.

Let's consider only models which describe outcomes that SB can consider on Sunday. This is a vague notion, that I hope to shore up with James... I think the claim is something like 'On Sunday, it may make no sense to consider the event "It is Monday"', as that changes during the course of the experiment. One response to this is to rule out events such as "It is Monday" in the probability model.

The only models that seem to be available then are models of the form:

$$P(\text{Heads}) = p,\ P(\text{Tails}) = 1 - p$$

Two popular such models are:

1. $P(\text{Heads}) = \frac{1}{2},\ P(\text{Tails}) = \frac{1}{2}$ (halfer)
2. $P(\text{Heads}) = \frac{1}{3},\ P(\text{Tails}) = \frac{2}{3}$ (thirder)

are any others possible?

Another 'thirder' probability model

Suppose that we aren't bothered by the idea that Monday/Tuesday change throughout the experiment, and on Sunday we use the probability model:

- $P(\text{Heads}, \text{Monday}) = \frac{1}{3}$
- $P(\text{Tails}, \text{Monday}) = \frac{1}{3}$
- $P(\text{Heads}, \text{Tuesday}) = 0$
- $P(\text{Tails}, \text{Tuesday}) = \frac{1}{3}$

This probability model is meant to reflect Sleeping Beauty's uncertainty on any given occasion that she is woken up. From a betting perspective it corresponds to bets that are guaranteed to be offered on any occasion that SB wakes up. James, can you express reasonably concretely what the problem with such models might be?