On the sleeping beauty "paradox"

Mark Holmes

February 27, 2023

Abstract

We discuss the sleeping beauty paradox or sleeping beauty problem, which has been a topic of discussion among philosophers, as well as in online forums such as youtube. The goal of this note is to describe the problem (or more precisely to point out that "the problem" is problematic), and to point out the elementary solution of some well-stated versions of the problem.

1 The experiment

The Sleeping Beauty problem or paradox seems to have first appeared in some form in the Philopshy literature [1], later in economics [**?**], and has achieved sufficient notoriety to appear on wikipedia [2] and various other online media. The problem concerns the following experiment, that Sleeping Beauty (SB) agrees to participate in. [1]

SB is magically put to sleep on Sunday. A fair coin is tossed. If the outcome is heads then SB is woken up only on Monday. If the outcome is tails then SB is woken up on Monday, put back to sleep, and woken up again on Tuesday. Whenever SB is woken up, she remembers that she is part of an experiment (and she remembers the rules of the experiment), but she is not told the outcome of the coin, nor what day it is, nor does she have any memory or physical indication of how long she has been in the experiment, nor whether she has been woken up before.

One can of course argue that such an experiment is not possible in practice as e.g. a human body/brain has physical memory that cannot be negated in such a way. The goal of this note is to discuss solutions to probabilistic questions that can be and are asked about this "theoretical" experiment.

[1] The experiment takes place in a land in which obtaining ethics approval is simple.

The question(s)

Modulo the impossibility of such an experiment from a practical point of view, and personal preferences for wording, grammar, etc., there is nothing wrong with the above description in the opinion of the author. The issues arise with what question(s) are then asked about the experiment. Here is an example appearing in the literature:

> When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads? [4]

Note that in this wording of the problem you (the reader) are the sleeping beauty. The choice of words here does not make it clear whether the question is asking for the probability of something, but since the paper goes on to calculate probabilities, it seems that it is intending to ask for the probability of something. There has been subsequent debate online about whether "the answer" should be 1/2 or 1/3. We will argue below that for any interpretation of the above as asking for the probability of something, the answer cannot be 1/3.

Note that it is possible to ask questions about this experiment that are not probabilistic. The author is not offering any commentary about answers to such questions, except to suggest that researchers interested in variants of this problem should take care to make their question explicit and as clear as possible.

We will also point out below that there are further structures that we can add to this problem (such as rewards, or repeated experiments), and questions that we can ask about those additional structures, for which 1/3 is a reasonable answer. However, we reiterate that this 1/3 is not the probability of something arising from the single coin toss that is described in the original problem.

The probability space

The only randomness apparent in the experiment is a single coin toss. The set of possible outcomes for this experiment is $\Omega = \{h, t\}$ and $\mathbb{P}(\{h\}) = \mathbb{P}(\{t\}) = 1/2$ assuming that the coin is a fair one. There are exactly 4 events in this problem: $\varnothing, H = \{h\}, T = \{t\}, \Omega = \{h, t\}$. Denote this collection by $\mathcal{F}$, i.e. $\mathcal{F} = \{\varnothing, H, T, \Omega\}$. Let $\mathcal{F}' = \{H, T, \Omega\}$. Any well-defined event for this problem must be one of the events in $\mathcal{F}$. For example, the event W that SB is ever woken up is in fact just Ω. The event that SB is woken up exactly once is H, and the event that SB is woken up exactly twice is T. The event that SB is woken up on a Tuesday is also T. The event that SB is woken up on a Monday is Ω. On the other hand "when she is woken up, it is Tuesday" is not an event.

There are exactly 12 quantities that one can calculate, of the form $\mathbb{P}(A|B)$ [2] for $A \in \mathcal{F}$

[2]By definition, $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ (where $\mathbb{P}(B) > 0$) and it refers to the probability that the outcome is one of the outcomes in A, given that it is one of the outcomes in B (this amounts to restricting the set of possible

$\mathbb{P}(\varnothing\|H) = 0$	$\mathbb{P}(\varnothing\|T) = 0$	$\mathbb{P}(\varnothing\|\Omega) = 0$
$\mathbb{P}(H\|H) = 1$	$\mathbb{P}(H\|T) = 0$	$\mathbb{P}(H\|\Omega) = 1/2$
$\mathbb{P}(T\|H) = 0$	$\mathbb{P}(T\|T) = 1$	$\mathbb{P}(T\|\Omega) = 1/2$
$\mathbb{P}(\Omega\|H) = 1$	$\mathbb{P}(\Omega\|T) = 1$	$\mathbb{P}(\Omega\|\Omega) = 1.$

Table 1: Table of all possible "probabilities" coming from the description of this experiment

and $B \in \mathcal{F}'$. See Table 1. If the question is asking for the probability of something (whether unconditional, or conditional on some event occuring), and it is well-posed then it should be clear which of the probabilities in Table 1 it is.

In particular, if you wish to ask "What is the probability that the coin toss is heads, given that SB is ever woken up?" Then you are asking for $\mathbb{P}(H|\Omega) = \mathbb{P}(H) = 1/2$. If you ask "What is the probability that the coin toss is heads, given that SB is woken up on a Monday?", the answer is the same, because you are asking for the same quantity. If you ask "What is the probability that the coin toss is heads, given that SB is woken up on a Tuesday?" then you are asking for $\mathbb{P}(H|T) = 0$. If you ask, "what is the probability that SB is woken up on a Tuesday, given that she is ever woken up?", then you are asking for $\mathbb{P}(T|\Omega) = \mathbb{P}(T) = 1/2$.

Note that none of the answers in the table above is $1/3$. Many people who have thought about the sleeping beauty problem get an answer of $1/3$ when calculating something. It is not clear what that something is, but since $1/3$ does not appear in Table 1 it is not a probability associated to the random experiment presented.

In the question from [4] stated in the previous section, the author of this note cannot tell which of the probabilities in Table 1 is being asked for (or in fact whether a probability is being asked for at all). So the author considers that the above question is not well-posed, and therefore any supposed solution to that question has no meaning. *If* the wording is intending to ask for $\mathbb{P}(H|W)$ then as noted above the solution is simple, as we have $W = \Omega$ so $\mathbb{P}(H|W) = \mathbb{P}(H|\Omega) = \mathbb{P}(H) = 1/2$.

Rewards and strategies for SB

Suppose that sleeping beauty is asked "Was the outcome of the coin heads or tails?" and the question posed to us is "What should she answer?". Then this is another ill-posed problem that cannot be answered without more information being provided. For example, one needs an objective function to answer this question: in particular, if it makes no difference to sleeping beauty how many correct or incorrect answers she gives in this experiment, then the answer to this particular question would be "it doesn't matter".

outcomes to B and rescaling the probabilities accordingly)

If instead she gains \$1 for each correct answer, and does not lose anything for an incorrect answer then she should decide in advance to always answer "tails" (assuming that she is allowed to remember her pre-experiment strategy when woken up). This is because if the outcome is h then she gains nothing under this strategy, while if the outcome is t then she gains \$2. If instead her strategy is to always answer "heads" then she gains \$1 if the outcome is h and nothing if the outcome is t. If she is allowed a randomized strategy (e.g. from rolling a die) that is independent of the coin toss of the experiment then she has potential gains no matter what the outcome of the coin toss of the experiment, but the distribution of her gain is stochastically dominated by the "always say tails" strategy above.

Perhaps this provides one possible way of understanding why some people get an answer of 1/3. She has the opportunity to be correct twice when the coin toss is tails, but only once when the coin toss is heads. So 2 out of 3 of the opportunities to give correct answers occur when the outcome is tails.

Again we note though that this 1/3 is not a probability appearing in Table 1. A way of viewing this argument from a probabilistic point of view occurs if we consider repeated experiments.

Repeated experiments

Suppose that the experiment is repeated many times. Let R_n be the number of coins that come up tails after n independent repetitions of this experiment. Let N_n be the number of times that SB has been woken up after n independent repetitions of this experiment. The quantities R_n and N_n are random variables whose value can be determined if we are told the outcomes of the first n coin tosses.

Then $N_n = 2R_n + (n - R_n) = R_n + n$ since SB is woken up twice for each tail, and once for each head. The law of large numbers in probability tells us that with probability 1, R_n/n converges to $\mathbb{P}(T) = 1/2$. It follows that $N_n/n \to 3/2$ with probability 1.

Suppose that SB has the strategy of always answering "tails" whenever she is woken up. Then she is correct exactly $2R_n$ times. The proportion of time that she is correct is therefore

$$\frac{2R_n}{N_n} = \frac{2R_n/n}{N_n/n} \to 2/3.$$

In other words, in the long run she will be correct 2/3 of the time if she always answers tails. Similarly she will be correct 1/3 of the time in the long run if she always answers heads.

This however is quite a different random experiment to the one described e.g. in [4]. Here we have infinitely many coin tosses, whereas the question that we started with has only one.

Acknowledgements

The author thanks Peter Taylor for making him aware of this problem.

References

[1] A. Zuboff, A. "One Self: The Logic of Experience". *Inquiry: An Interdisciplinary Journal of Philosophy.* 33 (1): 39–68, (1990).

[2] Sleeping Beauty Problem. `https://en.wikipedia.org/wiki/Sleeping_Beauty_problem`

[3] M. Piccione, A. Rubenstein, A. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior* 20: 3–24, (1997).

[4] A. Elga. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2): 143-147, (2000).