Sleeping beauty

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 ½ seems straightforward: Sleeping Beauty knows that the coin is fair, and so also knows that there is a ½ chance that it comes up heads on any given throw, and a ½ 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 ½ chance that the coin came up heads.

This involves some 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.

However, there are powerful arguments for the conclusion that Sleeping Beauty should **shift** the probability she assigns to the coin being heads from ½ before she is put to sleep, to ⅓ after she is awoken.

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.

Let’s adopt a few abbreviations:

T1: The coin came up tails and it is Day 1.
T2: The coin came up tails and it is Day 2.
H1: The coin came up tails and it is Day 1.

We’ll now need to introduce a little bit of notation for talking about the probabilities of these various states of affairs:

P(X): The probability of X
P(X I Y): The probability of X given Y; roughly, the probability you should assign to X, given the knowledge that Y is the case.

We can now argue as follows that we should shift our probability in the coin having come up heads to 1/3 upon being woken up:

1.	$(P(T1 \mid T1 \text{ or } T2) = (P(T2 \mid T1 \text{ or } T2)$	premise
2.	$P(T1) = P(T2)$	(1)
3.	$(P(H1 \mid H1 \text{ or } T1) = \frac{1}{2}$	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?

Let’s consider a second argument in favor of ⅓.

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.

Let’s consider a second argument in favor of ⅓.

We can imagine changing the case so that whether or not the coin comes up heads, Sleeping Beauty is awakened **both** days. In this version, when Sleeping Beauty is awakened, her probability assignments should clearly be as follows:

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

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? Well, 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 we can infer that

$$P(H1)=\tfrac{1}{3}.$$

But it seems that our modified case is the same as the original one: we can rule out, given the rules of the game, H2; and the other three possibilities seem equiprobable.

There's also a kind of intuitive argument for the conclusion that P(H1)=⅓. 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-awakenings. 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 ⅓. (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)=⅓?)

Nonetheless, it is hard to get rid of the intuition that when asked “What are the odds that a fair coin came up heads?”, one should always say: ½. **Especially given that upon being woken, we seem to learn nothing new — after all, we knew that we would be woken on one of the days.**

> 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.

Nonetheless, it is hard to get rid of the intuition that when asked “What are the odds that a fair coin came up heads?”, one should always say: ½. **Especially given that upon being woken, we seem to learn nothing new — after all, we knew that we would be woken on one of the days.**

Some indication that this is more than an intuition is given by consideration of the **generalized Sleeping Beauty problem**.

Imagine that we vary the original case as follows: each time you would be awoken in that scenario, you have a 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?

It seems quite plausible that you should reason as follows: first, the probability that you will be woken up once, given that heads was flipped, is pretty clearly 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:

$$1-(\tfrac{99}{100})^2$$

this is, intuitively, because the chances of you not being woken up on either day is the square of the chances of you not being woken up on one day (i.e., 99/100), and the chances of you being woken up once is 1 — the chances of you not being woken up on either day (since you will definitely either be woken at least once, or not woken either day, and not both).

So now imagine that you are woken up; you know then that you have been woken at least once. What are the odds that you should now assign to the coin having come up heads?

To figure out how to answer this question, an analogy may help: suppose that you think that the odds of ND winning the National Championship next year are 15/100, and the odds of Purdue doing so are 1/100. Suppose you are now told that next year the National Championship winner will come from the state of Indiana. (IU and Ball State have 0 chance.) What odds should you now assign to ND winning the National Championship?

A natural thought is: 15/16, since we have now eliminated 84 of the 100 “possibilities.” Slight more formally, it seems that we should take it to be the odds of ND winning before learning the information about Indiana over the odds of ND winning plus the odds of Purdue winning (again, prior to our information about the NC winner coming from Indiana).

Let’s apply this to the generalized Sleeping Beauty.

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.

Some indication that this is more than an intuition is given by consideration of the **generalized Sleeping Beauty problem**.

Imagine that we vary the original case as follows: each time you would be awoken in that scenario, you have a 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?

It seems quite plausible that you should reason as follows: first, the probability that you will be woken up once, given that heads was flipped, is pretty clearly 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:

$$1 - (\tfrac{99}{100})^2$$

this is, intuitively, because the chances of you not being woken up on either day is the square of the chances of you not being woken up on one day (i.e., 99/100), and the chances of you being woken up once is 1 — the chances of you not being woken up on either day (since you will definitely either be woken at least once, or not woken either day, and not both).

So now imagine that you are woken up; you know then that you have been woken at least once. What are the odds that you should now assign to the coin having come up heads?

Let’s apply this to the generalized Sleeping Beauty.

In this case, you know that the coin came up either heads or tails, and you were woken; so the probability that the coin came up heads is presumably the probability of you being woken given heads divided by the probability of you being woken given heads + the probability of you being woken given tails, i.e.:

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

So this says that the probability you should assign to heads = just over ⅓. The interesting thing, though, is that the probability of heads, using this reasoning, steadily increases as the odds of you being woken up on the various occasions increases.

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.

Some indication that this is more than an intuition is given by consideration of the **generalized Sleeping Beauty problem**.

Imagine that we vary the original case as follows: each time you would be awoken in that scenario, you have a 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?

In this case, you know that the coin came up either heads or tails, and you were woken; so the probability that the coin came up heads is presumably the probability of you being woken given heads divided by the probability of you being woken given heads + the probability of you being woken given tails, i.e.:

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

So this says that the probability you should assign to heads = just over ⅓. The interesting thing, though, is that the probability of heads, using this reasoning, steadily increases as the odds of you being woken up on the various occasions increases.

For example, suppose that there is a 99/100 chance, rather than a 1/100 chance, that you will be woken in the relevant occasions. Then, using the above reasoning, the probability which should be assigned to heads works out as follows:

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

And, in general, as the odds of being woken on the relevant occasions approaches 1, the probability which should be assigned to heads approaches ½ — which is strong evidence that the right response to the initial Sleeping Beauty problem is ½, not ⅓. If this is right, then something must be wrong with the three arguments for ⅓ which we have considered.

Or more, generally, we can say that **either** there is something wrong with what we have said about the generalized Sleeping Beauty, **or** there is something wrong with the arguments we have presented for ⅓. What you should think about is which of these you think is correct.

If the correct answer is ⅓ — and there is something wrong with the above argument about generalized Sleeping Beauty — there may be an interesting parallel between the two-envelope paradox and Sleeping Beauty.

If the correct answer is ⅓ — and there is something wrong with the above argument about generalized Sleeping Beauty — there may be an interesting parallel between the two-envelope paradox and Sleeping Beauty.

Remember that we were discussing the relationship between the “open” and “closed” versions of the paradox, and I suggested that we only get the truly paradoxical results which follow from consideration of the closed versions if we adopt the following principle:

Inference from an unknown

Suppose that you are choosing between two actions, act 1 and act 2. It is always rational to do act 2 if the following is the case: there is truth about the situation which you do not know but which is such that, were you to come to know it, it would be rational for you to do act 2.

I suggested that there is something of a puzzle about how this principle could be false, but that rejecting it seems to give us a decent treatment of the various versions of the two-envelope paradox we discussed. Sleeping beauty — if ⅓ is the right view — seems to be another case which leads us to reject this principle.

After all, if ⅓ is the right answer, it seems that we can describe Beauty’s situation as follows: before being put to sleep, she knows that she will be awoken; and she knows that upon learning that she has been awoken, she will be rational to judge that the odds of the coin having come up heads is ⅓. But now, prior to being woken, she is not rational to judge that the odds that the coin will come up heads is ⅓ — now, obviously, the right answer is ½, despite the fact that she knows that she is about to undergo an experience which will lead her to revise that estimate.

However, Sleeping Beauty offers a further challenge. Even if we reject inference from an unknown, what is the relevant unknown? One wants to say something like: the proposition that I have been woken **today**. But suppose that this is day 1; is this different than the proposition that I was woken on day 1 (which, after all, I already knew would be the case)?

The intuitive idea is that what I learn upon being awoken is that **now** is a time at which I am awake after the coin flip, and that this is a different claim than the claim that Day 1 is a time at which I am awake after the coin flip, even if Day 1 is now.

The sort of knowledge that Beauty gains when she is awoken is sometimes called “self-locating knowledge” — intuitively, it is not knowledge about what the world is like, but rather knowledge about where (and when) one is in that world.

The sort of knowledge that Beauty gains when she is awoken is sometimes called “self-locating knowledge” — intuitively, it is not knowledge about what the world is like, but rather knowledge about where (and when) one is in that world.

Other examples of that knowledge give rise to puzzles which are in some ways analogous to Sleeping Beauty. One is the paradox of the Eternal Coin, which is due to Cian Dorr.

Suppose that there is a fair coin which will be flipped every day in an infinite series of days — extending infinitely many days into the past, and infinitely many days into the future. Now let’s adopt the following abbreviations:

H = The eternal coin will come up heads today.

F = The eternal coin will come up heads every day in the future.

P = The eternal coin came up heads every day in the past.

Now let’s ask the question: what is the probability of H, given F?

The obvious answer is: ½. (Remember: you’re not supposed to doubt that the coin is fair when you learn that F is true; you’re still sure that it is a fair coin, just one that will, in a very unlikely coincidence, come up heads infinitely many times in a row, beginning tomorrow.)

But Dorr gives an argument that the correct answer is 1. This is a very surprising conclusion; let’s consider the argument. (We’ll consider a simplified version of one of three arguments that Dorr gives for this conclusion.)

Consider the following situation: On Sunday night you take a drug which you know will make you sleep, unless you’re awoken, until Thursday. But before you go to sleep I promise you that I will wake you up on both Monday and Tuesday. But, on each day, I will put you back to sleep after giving you a drug which will cause to forget having been woken. That’s all you know; nothing in your evidence suggests that I will, e.g., act differently when waking you on Monday as opposed to Tuesday. You go to sleep. Upon my waking you up, there’s no further evidence about what day it is — there are, e.g., no calendars in the room with big X’s on days that have passed. What probability should you assign to the claim that it is Monday?

The obvious answer appears to be: ½.

Now imagine that we vary the case by saying that I promise to wake you up on Monday, Tuesday, or Wednesday — and, as above, you have no evidence favoring one of those days over the others. In this case, what probability should you assign, upon being awoken, to the claim that it is Monday?

Now it seems that the answer is: ⅓. But, you might ask, where are these numbers coming from? What principle are we using to arrive at these judgments?

H = The eternal coin will come up heads today.

F = The eternal coin will come up heads every day in the future.

P = The eternal coin came up heads every day in the past.

Now it seems that the answer is: ⅓. But, you might ask, where are these numbers coming from? What principle are we using to arrive at these judgments?

It seems like we are making use of a kind of **indifference** principle, which says that, in absence of evidence favoring one hypothesis over another, one should assign them both equal probabilities.

But this is a special sort of indifference principle, since we are not really entertaining different hypotheses about what the world is like — you know, in the second case, that you will be awoken on Monday, Tuesday, and Wednesday. What you are doing is not considering different hypotheses about what the world is like — what you are doing is considering different hypotheses about where (really, when) in that world you are. Let's call this principle self-locating indifference:

Self-locating indifference

If A and B are inconsistent claims about where/when in the world you are, and your evidence does not count in favor of either, you should assign A and B equal probability. Hence, conditional on learning that A or B is true, Pr(A)=Pr(B).

Can you see why this principle would explain our intuitive judgements about the example of you being woken up, and asked which day it is?

Now let's return to the paradox of the Eternal Coin, and our question of what the probability of H, given P is.

H = The eternal coin will come up heads today.

F = The eternal coin will come up heads every day in the future.

P = The eternal coin came up heads every day in the past.

Self-locating indifference

If A and B are inconsistent claims about where/when in the world you are, and your evidence does not count in favor of either, you should assign A and B equal probability. Hence, conditional on learning that A or B is true, Pr(A)=Pr(B).

Now let's return to the paradox of the Eternal Coin, and our question of what the probability of H, given F is.

Now consider the following series of self-locating claims:

F1: The coin will conclude with an infinite series of heads, and tomorrow is day #1 of that series.

F2: The coin will conclude with an infinite series of heads, and tomorrow is day #2 of that series.

F3: The coin will conclude with an infinite series of heads, and tomorrow is day #3 of that series.

F4: The coin will conclude with an infinite series of heads, and tomorrow is day #4 of that series.

F5: The coin will conclude with an infinite series of heads, and tomorrow is day #5 of that series.

…..

Does your evidence favor any one of these over the others?

It seems not. But then upon learning that one of them is true, it follows by our Self-Locating Indifference principle that you should regard all of them as equiprobable.

But now recall our question of the probability of H given F. If you think about it, you will see that F is equivalent to the claim that F1 or F2 or F3 or …. Hence upon learning that F is true, you should assign all of F1, F2, F3, … equal probability.

But what does this tell us about H? Well, we know that if one of F1, F2, F3, … is true, **the only way for H to be false is for F1 to be true**. And what is the probability of F1? Well, we know that (given F) F1 is equiprobable with infinitely many other claims with which it is inconsistent. Hence, given F, Pr(F1)=1/∞ = 0 (or an infinitesimally small number). Hence the probability of not-H given F is 0. But we know that either H or not-H is true. Hence Pr(H I P)=1 (or a number infinitesimally close to 1).

This is very hard to believe. But the only way to reject it seems to be to reject Self-Locating Indifference, which seems quite plausible. One question you may want to ask yourself is: if we do reject Self-Locating Indifference, how does this affect our judgements about the Sleeping Beauty Paradox? Did we implicitly rely on a principle of this sort when assigning probabilities to the relevant claims?