The Sleeping Beauty Paradox Resolved

Abstract

The Sleeping Beauty problem is a famous open paradox about probabilities that has divided and polarized communities of mathematicians – probability theorists, decision theorists and even philosophers – for over fifteen years. This simply stated problem in self-locating belief, involves a rational agent undergoing amnesia (https://en.wikipedia.org/wiki/Sleeping_Beauty_problem). It appears like a seemingly simple puzzle in subjective probability. It has two possible solutions, one-half and one-third, both of which have staunch adherents, the "halfers" and the "thirders." Using frequentist methods, the answer clearly seems to be one-third (Elga 2000); yet, this equally clearly violates a basic natural Bayesian assumption that belief should not change in the absence of new evidence (Lewis, 2001). The passion generated in arguments between these two entrenched camps, which I recently experienced firsthand in a column for Quanta magazine (https://www.quantamagazine.org/20160114-sleeping-beautys-necker-cube-dilemma/) puts political ideological debates to shame. This has led several commenters to examine and question the validity of subjective probability or "credence." Yet, interestingly, competent halfers and thirders reach identical correct solutions when challenged with well-specified variations of the original problem, regardless of whether the solutions are consonant with their positions or not.

I believe I have pinpointed the reason for this passionate polarization. There are two alternative interpretations of this puzzle, hopelessly entangled, that are based on two different construals of the problem statement. These construals are based on two different, perfectly natural, ways that past events can be imagined. Each of the alternate ways, which I call the "action interpretation" and the "property interpretation" respectively, has strong intuitive appeal, and guides the way that different individuals think about and solve the problem. My resolution of the paradox shows that both Bayesian and frequentist methods give the same answers and demonstrate that there is no quarrel between the definitions of the notion of credence of the halfers and the thirders: It's just that the two camps apply their understanding to two completely different propositions. This is the reason why both camps are constantly talking past each other. This resolution of the paradox is consistent with that proposed by Berry Groisman (2008), giving two separate probabilistic experiments. My treatment pinpoints in a novel way, the exact place where ambiguity creeps in to cause the paradox, and shows how the naturalness of the two alternative interpretations completely explains the intuitive reasons for the passion and polarization that this paradox continues to generate.

Elga, A. (2000). "Self-locating Belief and the Sleeping Beauty Problem". Analysis 60 (2): 143–147. doi:10.1093/analys/60.2.143. JSTOR 3329167.

Lewis, D. (2001). "Sleeping Beauty: reply to Elga". Analysis 61 (3): 171–76. doi:10.1093/analys/61.3.171. JSTOR 3329230.

Groisman, B. (2008) "The end of Sleeping Beauty's nightmare" arXiv:0806.1316

The Sleeping Beauty Problem, Resolved!
Episode 2: Lost in Time

The Sleeping Beauty problem, described below, is a famous open problem about probabilities that has divided and polarized communities of mathematicians – probability theorists, decision theorists and even philosophers – for over fifteen years. The passion generated in arguments between the two entrenched camps, the "halfers" and the "thirders" puts political ideological debates to shame. In a recent Quanta Insights column, I compared the problem to a Necker cube, a famous visual illusion which can be perceived in two mutually exclusive ways. Most people can flip quite easily between the two views of the Necker cube, however, while in case of the Sleeping Beauty problem halfers and thirders remain firmly entrenched in their view, stubbornly experiencing the other view as completely wrong. Yet, curiously, as we saw in a previous column, competent halfers and thirders reach identical correct solutions when challenged with well-specified variations of the original problem, regardless of whether the solutions are consonant with their positions or not. Both camps can certainly do math, so what makes them butt heads so hard in vain? Is the problem underspecified? Is it ambiguous?

Here is the problem:

> The famous fairy-tale princess Sleeping Beauty participates in an experiment that starts on Sunday. She is told that she will be put to sleep, and while she is asleep a fair coin will be tossed that will determine how the experiment will proceed. If the coin comes up heads, she will be awakened on Monday, interviewed, and put back to sleep, but she won't remember this awakening. If the coin comes up tails, she will be awakened and interviewed on Monday and Tuesday, again without remembering either awakening. In either case, the experiment ends when she is awakened on Wednesday without being interviewed.
>
> Whenever Sleeping Beauty is awakened and interviewed, she won't know which day it is or whether she has been awakened before. During each awakening, she is asked: "What is your degree of certainty* that the coin landed heads?" What should her answer be?

*The phrase "degree of certainty" has been variously expressed as "belief", "degree of belief", "subjective certainty", "subjective probability" or "credence."

It seems very hard to believe that this simply stated problem should have remained open for over fifteen years. It is possible that the problem is underspecified, but it doesn't feel that way – both camps feel confident that they have solved it. This fact

hints at some deep ambiguity in the problem's statement about which smart people vehemently disagree. In our previous column, we saw some dichotomies in the two camps: Halfers count experiments, thirders count awakenings; halfers calculate from the experimenter's point of view, thirders from Sleeping Beauty's. But these are mathematical techniques that both camps know how and when to use. If they do reach different conclusions, it is probably not a matter of mistaken calculation: they must, in effect, be solving two entirely different problems.

This point – that the problem is ambiguous, and that both sides are correct – has been made by several commenters in web discussion groups. A paper by Berry Groisman showed that there are two interpretations of the problem that are both consistent under standard probability theory. I agree with this viewpoint, but it still remains to be explained why both halfers and thirders are so entrenched in their view, and are strongly convinced that they are right. This strong feeling arises, I suggest, because this quarrel is not about mathematics, but rather, it is actually a hidden, subconscious fight about *two different ways of understanding the problem statement.* Specifically, there are two equally valid interpretations or construals of the phrase "landed heads" which refers to something that happened in the past.

To reveal these two meanings, let me add a small hitherto unknown detail to the famous Sleeping Beauty story.

> Imagine that when the coin was tossed on Sunday, it was mounted on a brass plaque in the position it landed, so that the result of the coin toss can be checked at any time. This plaque is kept in a locked safe in Sleeping Beauty's room.

That's all we need to add to prime our intuitions. Certainly this act cannot, in any way, make any difference to the logic of the actual problem. But it enables us to see that the original question can be interpreted in two different ways:

> Interpretation 1: The Action Focus: "What is your degree of belief that the coin landed heads" = What is your degree of belief that the coin landed heads *in the act of tossing*?
> (Image to prime your intuition: Imagine the coin being tossed)
>
> Note that the belief, though current, is about a previous event: The verb "landed" is in the past tense. Whenever a past event is evoked in speech or writing, the listener or reader has to decide how much of the event's background is relevant. Sometimes, a phrase referring to the past requires the listener to "import" the event's background without the speaker explicitly saying so – a "frozen past tense."[1] It's like a photo taken when you were ten years old, unchanging forever,

[1] Note that, as a couple of language experts including Professor Steven Pinker have told me, this is not a phenomenon specific to a particular language. How much of the past background one needs to

which shows your old house in the background, even though you've changed a lot and the house is gone. Here's a reference to the past that requires this kind of implicit background importing: "*What is your belief that my friend the rock star spent a full year's pay on his first guitar?*" This question refers to your belief about the money my friend was making when he bought the guitar, not what he makes today. After all, he is a rock star now, twenty years later. In a similar way, the first meaning of the Sleeping Beauty proposition imports its background act: It can be intuitively accessed by invoking the image of the coin being tossed. It refers to the probability that the coin landed heads when it was tossed: obviously one-half.

Interpretation 2: Property Focus: "What is your degree of belief that the coin landed heads" = What is your degree of belief that the preserved coin in the safe is *showing heads now*?
(Image to prime your intuition: Imagine what face is showing on the coin in the safe.)

Here the past tense predicate "landed heads" is used as a way to describe a *property* previously gained by the coin you are referring to today. You are no longer concerned with how likely it was to have landed heads or tails when originally tossed, but are now concerned with the likelihood that the coin being referred to today, the preserved coin, shows heads and therefore "landed heads" at some time in the past (as opposed to landed tails). The same past tense verb construction is used but now it does not import the background action.

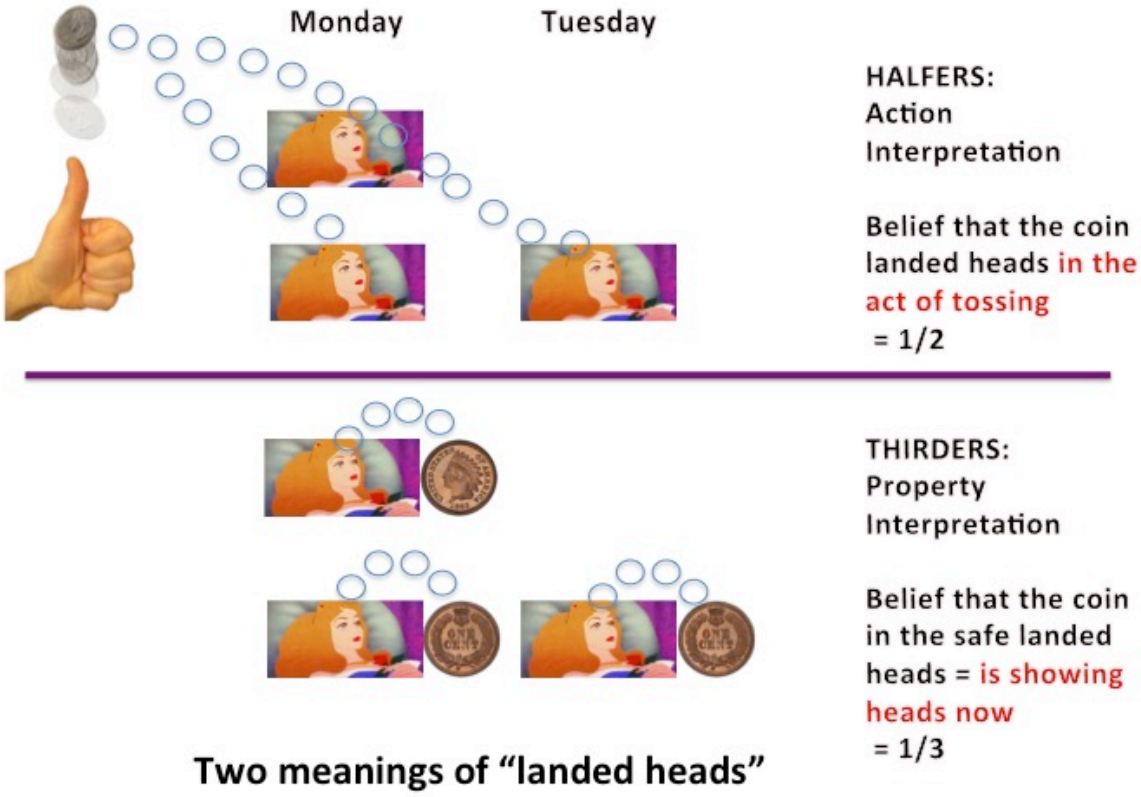
As an example, remember that in the year 2015[*], 6 of the 9 US Supreme Court justices attended Harvard Law School and 3 attended Yale Law School. If I had asked you then, *"What is your degree of belief that a random US Supreme Court Justice attended Yale Law?"* you would probably have answered 1/3. Here, the question is not intended to import the background of the event: I was not asking you about the odds that the person chose Yale Law School of all the law school choices they had back then, you are using "attended Yale" as a property of the person within the current group of Justices, as opposed to the opposite property, "attended Harvard." In a similar way, the second meaning of the Sleeping Beauty proposition does not import the past action, but merely looks at heads as a present property, previously acquired. It can be intuitively accessed by invoking the image of comparing the coin encountered with the plaque in the safe. It refers to the probability that the coin is showing heads now if you had to bet on it, or carried out this comparison on a number of similar occasions: this probability could be anything from 0 to 1. In this case it is 1/3.

---

assume for a given statement, is a universal problem when referring to the past, usually resolved by context.

[*]The numbers in this example have been rendered obsolete by the passing of Justice Scalia. I have kept the original numbers because of the memorable, and relevant ratio they generate.

In the real life example sentences I gave, it is quite clear from the context what is expected: Whether we should invoke the background of the event (the friend's previous salary) or merely use the acquired property (the justice's law school affiliation). However in case of a coin toss, the context does not force one interpretation or the other. Both are up for grabs – and boy, are they grabbed tightly by the two different camps!



Two meanings of "landed heads"

Halfers, I suggest, consciously or subconsciously find the first interpretation, image or intuition to be more salient, use it in their modeling, and come up with a value of one-half; while thirders subconsciously prefer the second interpretation, image or intuition, base their calculations upon it, and come up with a value of one-third.

How can this be? Isn't the coin in the safe the same coin that landed heads or tails on Sunday?

Let's ask the sophisticated and intelligent princess Sleeping Beauty, who is well versed in the natural arts and sciences such as linguistics, math and science, besides of course, fauna and flora. Let's catch her at the time of her interview with the experimenter's assistant at one of her awakenings.

EA: 'What is your degree of belief for the proposition that the coin landed heads?'
SB: 'That question is ambiguous: It can be interpreted in two different ways. Do you mean my belief about the likelihood of heads in the act of tossing the coin on Sunday or do you mean my belief about the likelihood of heads being shown on the preserved coin in the safe?'
EA: 'But the coin in the safe is the same coin that was tossed on Sunday, and shows the same result.'
SB: 'Yes, but you can have a different credence about the probability of heads in the act of tossing a fair coin (which is always one-half) and the probability of the heads in the same coin some time later.'

'Let's suppose that while walking on the seashore, I see 15 coins, 10 of which show tails and 5 heads. Perhaps, a boy who was on the beach before me, removed half of the coins that came up heads because he liked that side. No matter how many coins I gather, I always find two showing tails to every one showing heads. Half of the coins are, to me, **lost in space**. Now my credence that a new coin I encounter shows heads and therefore "landed heads" at some time in the past is only 5 out of 15 or one in three. I can only base my credence on the clear-cut and reliable statistics of the coins I encounter. Maybe one day, I'll find the boy's stash of coins that landed heads, and if I do, my expectation that there are equal numbers of coins that landed heads will return to one-half—or maybe I never will. Notice that we use the verbs "landed" or "came up" in two senses: in the act of tossing the coin, as in "the coin just landed heads," and in the act of finding it later as in "here's a coin that shows heads, and therefore landed heads sometime in the past." '

'Here's a different situation. Imagine I have a specific kind of double vision: the only objects it affects are coins showing tails. When a coin shows heads I see it as one; when it shows tails, a strange optical effect makes me see double. I actually saw only 10 coins, 5 showing heads and 5 showing tails. But my strange affliction, unknown to me, causes me to see heads and tails in the ratio 1:2 and hence my expectation that any new coin I find will show heads, and therefore landed heads, is one in three. Later if I find out that I have this condition, I can correct my erroneous, but at that time fully valid, belief. Of course, knowing that these are all fair coins, I never waver in my belief that they originally landed heads one in two times in the act of tossing.'

'Now, imagine that I find the same 15 coins, but I entered a time warp, unbeknownst to me, and five of the tails I saw were the same ones I had seen before. All the coins showing heads somehow escaped the time warp. Now half of the heads are **lost in time,** or alternatively, the tails are doubled in time. Again, my credence for heads is, validly, one-third. If and when I come out of the time warp, and realize it, I change my credence back to one-half.'

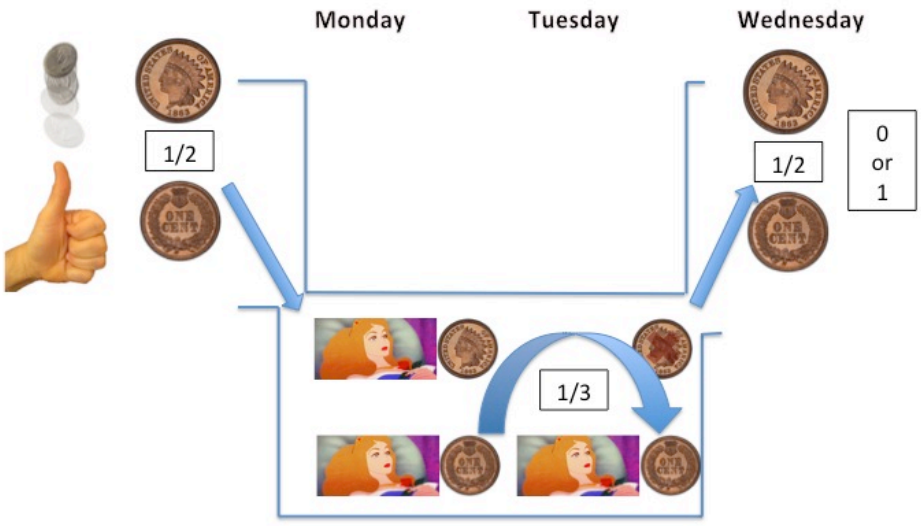
‘I don’t want you to think that such distortions necessarily reflect some kinds of errors of which I am unaware. It may be that half of all coins that landed heads self-destructed on landing so the reliable ratio of coins that I find actually reflects the existing ratio in the world. As long as I am unaware of any systematic errors, I therefore have to trust the ratio that I find, actually or by calculation, as the basis for my belief in the actual probability of the coins I am likely to find.’

‘Thus distortions in time, space and perception that I am unaware of, or differential longevities of the two kinds of coins, or *any systematic process that alters the frequencies* of the two coin toss results differentially, can alter the relative frequencies I systematically find. All these processes influence my valid belief regarding the proportion of the coins I am likely to encounter now that landed heads.’

‘Let’s return to your question, which is actually two separate questions.’

‘The first question is: What is my degree of belief that the coin landed heads *in the act of tossing*? This value, of course, was one-half on Sunday, and will remain one-half until I actually find out what happened. I am a true ***halfer*** about this.’

‘What is my degree of belief that the preserved coin is *showing heads now* which is the same as saying **“**what is my credence that the coin in the safe landed heads sometime in the past?” On Monday and Tuesday, it is one-third, because I am in a time warp with half the heads being lost in time. I am definitely a **thirder** on these two days.’



Heads lost in time:
Sleeping Beauty's experimental time warp

'When I will emerge from the experiment's time warp on Wednesday, the value of my credence for heads will once again return to one-half, *because the two different interpretations will coincide*. Then I expect that your boss, the Professor, will tell me how the coin actually landed. At that time, my credence that the coin landed heads will settle on exactly 0 or exactly 1.'

EA: 'Wow, you are one formidable princess. All that sleeping must be good for the brain! Ouch, my head hurts. I think I'll take some of the amnesia drug I will soon be giving you…'

That's all there is to it, halfers and thirders. The rest, as they say, is just plain… math.