Temporal Logic and Selection in the Sleeping Beauty Problem

Marc A. Burock

ABSTRACT

The Sleeping Beauty Problem is a polarizing thought experiment involving a fair coin toss, memory erasure and temporal uncertainty. Despite its simplicity there is no agreed upon solution. In this work I put forward a set of arguments that support the so-called *Halfer* or 1/2 solution to the problem, while undermining the competing *Thirder* or 1/3 solution. In analyzing Elga's original argument for the 1/3 solution, I bring to light a subtle but clear contradiction in his reasoning using temporal logic. Temporal reasoning also helps to neutralize the main criticisms against the 1/2 solution. Surprisingly, for some questions of probability or credence, it appears we need to distinguish between an event that has yet to occur, and the same event after it has already occurred. Knowledge that an event has been decided (without knowing the result) can be a type of admissible evidence when updating credences.

1 Introduction

The so-called *Sleeping Beauty Problem* is a simple thought experiment involving sleep and memory erasure, made infamous by Adam Elga (2000) and David Lewis (2001). It goes like this. A researcher is going to put you to sleep on Sunday and then toss a fair coin. If Heads then the researcher will wake

you up on Monday, ask you a question, then erase your memory and put you back to sleep until Wednesday. If Tails, the researcher will wake you up on Monday and Tuesday, both times asking you the same question, erase your memory, and put you back to sleep. You will wake on Wednesday for the completion of the experiment with no memory of what happened. Here is the question asked during your awakenings: what is your degree of belief (credence) that the coin toss was Heads?

We can imagine this experiment taking place in the actual world, and anyone with an introductory understanding of probability can take a crack at it. Yet starting with seemingly straight-forward premises, an analysis of the problem leads to two conflicting solutions—a degree of belief in Heads of 1/3 (*Thirders*) or 1/2 (*Halfers*). Try for yourself a solution if you have not yet taken a side in the dispute. Despite an impressive literature on Sleeping Beauty by academic scholars and hobbyists, the controversy continues.

Before I continue with the meat of the argument, I would like to briefly focus on Wednesday, or the third and last day of the experiment, as it is sometimes neglected in the literature. At the end of the experiment, the subject (I will refer to Sleeping Beauty as 'the subject' or 'you') must wae and stay woke, and the researcher will inform him that the experiment is over. After finishing the experiment, the subject does not know how many awakenings occurred prior to Wednesday, and the researcher does not divulge this information. Nonetheless, the researcher may ask the subject, as he is leaving the lab, what his credence in Heads is for the completed experiment. From the subject's perspective, the last thing he recalls is being put to sleep prior to beginning the experiment, and now only learns that the experiment is over. Prior to being put to sleep, the subject's (and researcher's) credence in Heads was 1/2. After the experiment, the subject knows his temporal location and that he participated in the experiment, but nothing about the specifics of that participation, just as prior to the experiment he knows his temporal location and that he will participate but nothing more specific. He has no reason to believe that the coin is unfair on Wednesday. The subject's epistemic state with regard to the experiment, prior to and after the experiment, are symmetric in this sense; so the subject's credence in Heads will be 1/2 afterward.

The subject's situation during a Monday or Tuesday awakening is different. Upon awakening, she randomly enters the world in the sense that she is uncertain about the day of the week, uncertain whether the coin landed Heads or Tails, and uncertain whether she will wake a total of one or two times. Prior to the experiment, the subject contemplates this uncertainty, and reasons about what her credence in Heads ought to be when she awakens in this uncertain state, knowing everything I told you about how the experiment is performed.

In the following section I will present arguments that support the 1/2 solution to the Sleeping Beauty problem while undermining the 1/3 solution. I don't suspect that any of these arguments taken independently will sway the opinion of a die-hard Thirder. I myself oscillated between the Halfer and Thirder positions for some time before I fell into the Halfer position securely. In section 3 I highlight a subtle probabilistic contradiction in Elga's original argument for 1/3 that has not been addressed in the literature before, and resolve the contradiction by applying a temporal logic to the coin toss. For those who stick around, in Appendix B I derive a single expression for a Repeated Sleepy Beauty experiment that unites Halfers and Thirders in the long run.

2 Arguments for 1/2

2.1 The total number of awakenings argument

Thirders and Halfers distribute uncertainty across awakenings differently (Table 1). Everyone agrees there are three possibilities, but may disagree on the credence for each.

Label	Situation	Thirders Credence	Halfers Credence
H1	HEADS and it is Monday	1/3	1/2
T1	TAILS and it is Monday	1/3	1/4
T2	TAILS and it is Tuesday	1/3	1/4

Table 1. *Credences for centered awakening events*

Suppose we ask the subject, upon awakening, "What's your credence of waking up two times total during the experiment?" The Thirder distribution implies that the subject's credence for waking up a total of two times during the experiment is 2/3, but this credence is incorrect according to a naive frequentist interpretation of probability: if we perform the experiment a large number of times, about 1/2 of those trials will have a total of two awakenings, not 2/3 of them. While the Thirder may argue that the credence of a Tails-awakening is 2/3—because there are twice as many Tails-awakenings as Heads-awakenings in the long run—this argument cannot be used on the *total* number of awakening per experiment. Yet landing Tails and waking up two times total ought to have the same credence.

Frequentist arguments famously do not settle the problem when only considering the side of the coin because we can use two different reference classes to calculate the long run frequency of Tails(or Heads). We might consider the number of Tails-awakenings or the number of times Tails occurred during repeated experiments, which give conflicting answers. But when we focus upon the total

number of awakenings per experimental run, there is only one way to carve up the reference class, and the Thirder gives the incorrect credence, according to a frequentist interpretation (as well as from a perspective of objective chance.)

There is, however, a related situation where your credence in waking up two times total is 2/3. If we imagine that the Sleeping Beauty experiment is repeated a large number of times, and I ask you to choose a single awakening 'at random' from the collection of all awakenings across experimental trials, and then ask: "What's your credence that you wake a total of two times in the experimental run *containing the awakening you chose at random*?", then the correct credence is 2/3. You are more likely to choose an awakening associated with experimental runs that have more awakenings—this is the so-called problem of 'random incidence'. But this is a different question. Rather, we want to know your credence in waking two times for this run, and not in some run associated with a randomly selected awakening across repeated experiments.

Here is another example to help see the difference. Imagine you have two cookies. One cookie has only one chocolate chip on it and the other has two chocolate chips on it. Have someone flip a fair coin and put the one-chip cookie in the bag if Heads *or* the two-chip cookie in the bag if Tails (but not both at the same time). Without knowing which cookie is in the bag, randomly select a single *chip* from the bag—someone or something can do this for you by giving you the only chip if it's a one-chip cookie, or by picking between the two chips with a fair coin if it is a two-chip cookie. What is your credence in getting a chip from a cookie with two chips? Let us call this the One-Cookie-In-A-Bag-Experiment (OCIABE). Next perform the Two-Cookies-In-A-Bag-Experiment (TCIABE) by putting both the one-chip and two-chip cookies in a bag. Have someone or something randomly select a *chip* from the bag by performing an equally likely selection between the three chips, and then give you the chosen chip. Again, what is your credence in receiving a chip from the two-chip cookie?

Your credence in getting a chip from the two-chip cookie in the OCIABE is 1/2, while this credence should be 2/3 in the TCIABE. Although for each experiment you will receive one of three possible chips, the experiments differ by mechanism of selection of those chips. In the cookie-chip experiment with two cookies in the bag, a single mechanism selects among the three chips contemporaneously, and you are more likely to receive a chip from a two-chip cookie—because it has more chips and selection was equally likely. For the one-cookie-in-a-bag experiment, no single mechanism selects between the three possibly chips contemporaneously. A mechanism first selects the one or two-chip cookie, and then another mechanism is either choosing the one and only chip, or selecting one chip on a two-chip cookie to give you.

While neither cookie experiment is structurally identical to the Sleeping Beauty experiment, the OCIABE copies it fairly closely for one selection, minus episodes of amnesia. Chips correspond to specific awakening events, the one-chip cookie corresponds to one awakening total when Heads, and the two-chip cookie to two awakenings total when Tails. Like chips in the OCIABE, selection of awakenings in the Sleeping Beauty experiment occur in two distinct stages—first the fair coin selects for one or two total awakenings, then the researcher 'selects' (causally initiates) awakenings to give the subject, dependent upon the first stage of selection. While all three awakenings are epistemic possibilities for a single run of the Sleeping Beauty experiment, they are not contemporaneously selectable within a single mechanism or act of selection.

The two situations can be expressed by subtly different samples spaces in set notation: $\Omega_1 = \{H1, T1, T2\}$ vs $\Omega_2 = \{\{H1\}, \{T1, T2\}\}$, where Ω_2 respects the two-stage nested selection of the Sleeping Beauty experiment and Ω_1 does not. The sample space Ω_2 further expresses the fact that *T1* and *T2* are mutually necessary events during the experiment—*T1* and *T2* occur together during the experiment or neither occurs—and further that $\{H1\}$ and $\{T1, T2\}$ are elements of selection. When we envision a mechanism that selects among the three awakening possibilities with one act of selection, similar to the TCIABE with chips, we distort the causality of the original Sleeping Beauty experiment. This is why Sleeping Beauty arguments which posit multiple runs of the experiment or multiple copies of the subject, and then allow for selection of a single awakening from this collection of awakenings are inappropriate models of the original experiment. In Appendix B I analyze a repeated Sleeping Beauty experiment and show how your credence in Heads can be 1/2 for one run of the experiment, and also approach 1/3 as the number of repetitions approaches infinity. This result also demonstrates that a credence derived from numerous repetitions may not apply to the single case.

2.2 The Mirror (Dual-information) Argument

Here is a direct argument for **P**(HEADS and it is Monday)=1/2. The researcher watching the subject undergo the sleeping beauty experiment can also assign credences to the various awakening outcomes (Table 2), where these awakening events will be temporally uncentered.

Situation	Researcher Credence
HEADS and subject wakes on Monday	1/2
TAILS and subject wakes on Monday	1/2
TAILS and subject wakes on Tuesday	1/2

Table 2. *Temporally uncentered awakening events*

The credences do not add to one because the Monday and Tuesday Tails awakenings are not mutually exclusive from the uncentered perspective, and the three events in Table 2 do not form a proper sample space. Suppose you (as the subject) had access to the researcher's perspective while you were undergoing the experiment. Define the following events:

HEADS=the fair coin lands on Heads

M_{UH} = HEADS and you wake on Monday (temporally uncentered)

M_{CH} = HEADS and it is Monday (temporally centered)

Then:

(**M1**) $\mathbf{P}(M_{UH})=\mathbf{P}(HEADS)=1/2$

(**M2**) $\mathbf{P}(M_{CH}|M_{UH})=1$, or in words, the credence this is a Monday awakening and Heads, given you wake on Monday and Heads, is one.

(**M3**) $\mathbf{P}(M_{UH}|M_{CH})=1$, the credence you wake on Monday and Heads, given this is a Monday awakening and Heads, is one.

(**M4**) $\therefore \mathbf{P}(M_{UH})=\mathbf{P}(M_{CH})=\mathbf{P}(HEADS)=1/2$, where $\mathbf{P}(M_{UH})=\mathbf{P}(M_{CH})$ follows from (**M2**), (**M3**), and the definition of conditional probability.

The Mirror Argument argues that our temporally centered and uncentered credences in "Monday and Heads" ought to be the same. When coupled to (**M1**) — which seems relatively uncontroversial — it gives **P**(HEADS and it is Monday)=1/2. If you deny (**M1**), but assume that $\mathbf{P}(M_{CH})=1/3$ as do Thirders, then the Mirror Argument requires that your uncentered credence in Heads and Monday is 1/3 as well. In other words, it requires the researcher who does not undergo an awakening to assign 1/3 to Heads! Less you think I surreptitiously smuggled in an identity to make this argument, consider a similar situation with Tails:

M_{UT} = TAILS and you wake on Monday (temporally uncentered)

M_{CT} = TAILS and it is Monday (temporally centered)

Then:

(**T1**) **P**(M_{UT})=P(*TAILS*)=1/2

(**T2**) **P**(M_{UT}|M_{CT})=1, the credence you wake on Monday and Tails, given this is a Monday awakening and Tails, is one.

(**T3**) **P**(M_{CT}|M_{UT})<1, the credence this is a Monday awakening and Tails, given you wake on Monday and Tails, is less than 1.

Why (**T3**)? Because given uncentered evidence that you wake on Monday and Tails, it is still possible that your current awakening is Tuesday and Tails. Your credence in M_{UT} is not automatically identical to your credence in M_{CT}. If we further apply indifference to these two outcomes, then **P**(M_{CT}|M_{UT})=1/2, and by Bayes Theorem **P**(M_{UT})=2·**P**(M_{CT}). Noting (**T1**), this would yield **P**(M_{CT})=1/4.

2.3 The Really Direct Argument

Suppose that the coin was tossed on Sunday, *prior to putting the subject asleep*. The subject will only wake on Monday if Heads, and on Monday and Tuesday if Tails. The subject's credence that it is Monday and Heads upon awakening is given directly by the definition of conditional probability:

P(M_{CH})=**P**(It is Monday|HEADS on Sunday)**P**(HEADS on Sunday)
=(1)·(1/2)

This first term follows from the structure of experiment. The second term is your uncentered, unconditional credence that a fair coin, tossed on Sunday, landed Heads.

2.4 The Random Selection Argument

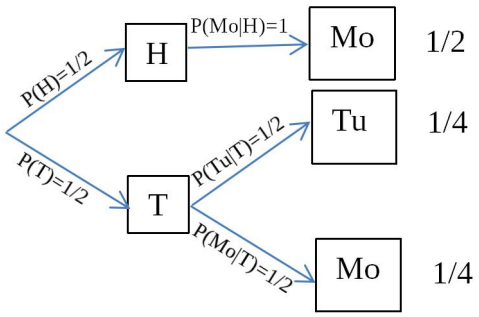

Figure 1. *Random selection model*

This argument is not new to the literature, but it does represent a new perspective and segues into what I think is the primary source of confusion in the Sleeping Beauty problem. Suppose that the coin has *already been flipped* and then you are awakened. You are uncertain about the side of the coin, and

uncertain about it being Monday or Tuesday. From your perspective, upon awakening, you randomly enter a world that is either Heads or Tails, and either Monday or Tuesday. You can reasonably model this situation as a random selection process in two steps: first randomly select a side of the coin, then dependent upon the side chosen, randomly select a day (Figure 1). Randomness can be interpreted as indifference in the coin's side, and then conditional indifference in the day given Tails. Two-stage selection mirrors the causality of the experiment. The structure of the experiment also demands that the day is Monday if the coin is Heads.

A common criticism about this distribution of credences is that **P**(*HEADS*|*MONDAY*)=2/3—why would being told it is Monday raise your credence in Heads? This would be a strange credence if it was Monday *and the coin had yet to be tossed*; in that case you should believe the credence in Heads is 1/2 less you violate the Principal Principle. But if the coin was already tossed, then given the random selection model above, 2/3 is precisely the correct credence in Heads given that it is Monday, and the Principal Principle is not violated because we are dealing with epistemic uncertainty in an outcome after a chance event already occurred.

It is easy, and consistent, to be a so-called double Halfer (hold a credence of 1/2 for Heads prior to flipping the coin), or triple Halfer (assign a credence of 1/2 to Heads after the experiment is over, too) with this model. Prior to flipping the coin, we do not invoke the random selection model—this model only applies during the experiment, after the coin has been flipped, on Monday and Tuesday. **P**(*HEADS*) is simply 1/2 prior to flipping the coin, independent of the day, in accordance with the setup of the example. After the experiment is over, I can likewise disregard the random selection model. The outcome of the coin does not influence what happens on Wednesday, or Thursday, or anything after the experiment is over. On Wednesday I am epistemically indifferent to Heads and Tails.

3 Temporal logic and a latent contradiction

I have suggested that whether the coin has already been tossed or not influences our reasoning. Along this line, Elga and those after him suggest two ways that the researcher may perform the Sleeping Beauty experiment:

1. *first* tossing the coin and then waking you up either once or twice depending on the outcome; or
2. first waking you up once, and *then* tossing the coin to determine whether to wake you up a second time.

Elga says that, upon awakening, the subject's credence in heads ought to be the same regardless whether method 1 or 2 is used to do the experiment, and proceeds to reason assuming that the researchers use method 2. Lewis agrees that the experimental method ought not change the result, yet the method of the experiment does influence the process of our reasoning, and Elga explicitly makes use of method 2 to make his argument. Specifically, he considers waking the subject on Monday prior to tossing the coin, then telling the subject that it is in fact Monday, and argues reasonably that **P**(*HEADS*|*MONDAY*)=1/2, yet also argues that **P**(*HEADS and it is Monday*)=1/3 on Monday, prior to being told it is Monday, and prior to tossing the coin.

I think method 2 generates a contradiction under probabilistic temporal analysis. To demonstrate this, I suggest making a distinction between an event that has yet to occur, and the same event after it has already occurred. In the case of the Sleeping Beauty problem:

H+	It will be HEADS in the future
-*H*	It was HEADS in the past
Mo	It is Monday now

I further assume a temporal symmetry principle that requires **P**(*H*+)=**P**(-*H*). With no other conditional evidence, your past credence that an event will occur in the future should be the same as your future credence that the same event already occurred in the past. Put slightly more formally: given chancy event *A*, let t_A be the time at which *A* is decided, t_{A-} is some time before t_A, and t_{A+} is some time after t_A . If you acquire no other evidence about *A* between t_{A-} and t_{A+} other than the knowledge that *A* has been decided, then your credence at t_{A-} that event *A* will occur in the future should be the same as your credence at t_{A+} that *A* occurred in the past.

If you recall the introduction to this paper, I loosely argued that your credence in Heads before starting the experiment should be the same as your credence in Heads after waking up and finishing the experiment on Wednesday, assuming the researcher did not leak any information to you. You knew the coin was fair before the experiment, and it should remain a fair coin after the experiment. The temporal symmetry principle above underlies the equality of credences. One might further argue that temporal symmetry gives the answer to the Sleeping Beauty Problem directly similar to Lewis (2001).

Nonetheless, although **P**(*H*+)=**P**(-*H*), *H*+ is not equivalent to -*H*, and I will show that the distinction can make a difference. In some cases, knowledge that an event has occurred can change your credence in the event. Consider this example: suppose there is a light that is red, it is now on, and you are about to flip a fair coin. If the coin lands Heads the light will change from red to green, but if

the coin lands Tails the light will stay red. Your credence **P**(*H+*|*Red*) should be equal to 1/2—it is a fair coin yet to be tossed, and the color of the light ought not change that. The situation for **P**(*-H*|*Red*) differs. If the light is red after the toss, then I know that the coin landed on Tails and that **P**(*-H*|*Red*)=0. Therefore **P**(*H+*|*Red*)≠**P**(*-H*|*Red*) and the temporal symmetry principle does not apply in this conditional setting. Knowledge that an event occurred can be admissible evidence when updating a conditional credence, and does not require a mysterious violation of the Principal Principle. More, it would be an error to assume that **P**(*H+*|*Red*)=**P**(*-H*|*Red*) and perhaps lead to a contradiction if *H+* and *-H* were not distinguished. With regard to the Sleeping Beauty problem, Elga's original argument for 1/3 also involves a situation where the occurrence of an event ought to count as admissible evidence, and although he does not explicitly distinguish the tense of Heads, he necessarily relies upon future-tensed reasoning:

> Your credence that you are in H1 would then be your credence that a fair coin, *soon to be tossed, will land Heads*. It is irrelevant that you will be awakened on the following day if and only if the coin lands Tails — in this circumstance, your credence that the coin *will land Heads* ought to be 1/2. But your credence that the coin *will land Heads* (after learning that it is Monday) ought to be the same as the conditional credence P(H1|H1orT1). So P(H1|H1orT1)=1/2, and hence P(H1)=P(T1).

In this case, Elga relies upon *H+* to make the argument, and does not make an appeal to *-H* nor suggest that *H+* and *-H* may differ. Therefore his argument entails:

(**E1**) **P**(*H+*)=1/2

(**E2**) **P**(*H+*|*Mo*)=**P**(*H+*)=1/2 , by objective chance and the Principal Principle

(**E3**) ∴**P**(*Mo*|*H+*)=**P**(*Mo*) , from (**E2**) since *Mo* and *H+* are independent

But (**E3**) creates a problem. The structure of the Sleeping Beauty experiment implies that **P**(*Mo*|*H+*)=1, yet this with (**E3**) implies that **P**(*Mo*)=1, which seems incorrect because the subject can wake on Tuesday as well as Monday. So **P**(*Mo*) must be less than one, but then this implies that **P**(*Mo*|*H+*)<1, which violates the structure of the experiment. Elga's method 2 of doing the experiment creates a paradox under analysis. Here is another way to see the contradiction when ignoring tense. Before the toss on Monday, Heads and Monday are independent of one another if you believe that **P**(*H*|*Mo*)=**P**(*H*)=1/2. After the toss 'on Monday night', then Heads and Monday become dependent

upon one another if you believe that $\mathbf{P}(Mo|H)\neq\mathbf{P}(Mo)$. But it violates probability theory to believe that both $\mathbf{P}(H|Mo)=\mathbf{P}(H)$ and $\mathbf{P}(Mo|H)\neq\mathbf{P}(Mo)$ are true. This is why we must distinguish between $H+$ and $-H$, for even though $\mathbf{P}(H+)=\mathbf{P}(-H)$, $-H$ may have developed future dependencies which do not exists for $H+$.

Prior to tossing the coin, $\mathbf{P}(H+)$ is anchored by the Principal Principle and objective chance, and $H+$ is independent of all non-exotic evidence. I could similarly argue that the statement of the Sleeping Beauty problem implies that the coin toss is unconditionally fair. But after the toss, the outcome of the coin is no longer a chancy event. Our uncertainty after the toss is epistemic uncertainty in an outcome that already occurred. If the particular outcome of the coin causes other events to transpire or is correlated with other events, and we acquire knowledge of those events after the toss, then we can conditionalize on that evidence to modify our credence in Heads *after the toss*. When the toss occurs on Monday night using method 2, then the toss is both independent of and dependent upon Monday, and we must make a distinction to avoid the contradiction.

The Random Selection Argument—which is a modified version of Lewis's original halfer argument—does not fall to the above contradiction because the reasoning is all done after the coin is tossed:

(**R1**) $\mathbf{P}(-H)=1/2$

(**R2**) $\mathbf{P}(-H|Mo)=2/3$, by the Random Selection Argument

(**R3**) $\mathbf{P}(Mo|-H)=1$, by the structure of the Sleeping Beauty problem

Since (**R1**) does not equal (**R2**), $-H$ and Mo are dependent, and it follows that $\mathbf{P}(Mo|-H)$ need not equal $\mathbf{P}(Mo)$, thus avoiding the paradox as well as a violation of the Principal Principle. It is interesting—but not inconsistent—that $\mathbf{P}(H+|Mo)\neq\mathbf{P}(-H|Mo)$, which is similar to the example with the red and green lights above. Before the coin is tossed the side it will land on is independent of the day, but after the toss, whether it landed Heads or not is dependent on the day. The temporal symmetry principle does not apply. Elga's argument implicitly demands that $\mathbf{P}(H+|Mo)=\mathbf{P}(-H|Mo)$, but this assumption leads to a contradiction. The above expressions lead to two different but direct formula for 'Heads and It is Monday', depending upon whether the coin was already tossed or not:

(**E4**) $\mathbf{P}(H+ \,\&\, Mo)=\mathbf{P}(H+|Mo)\mathbf{P}(Mo)=1/2\cdot\mathbf{P}(Mo)$

(**R4**) $\mathbf{P}(-H \,\&\, Mo)=\mathbf{P}(Mo|-H)\mathbf{P}(-H)=1/2$

The expressions **P**(*H*+|*Mo*), **P**(*Mo*|-*H*), and **P**(-*H*) in (**E4**) and (**R4**) were chosen because each is relatively uncontroversial. Our credence in Heads and Monday depends upon whether the coin toss has already occurred or not. Specifically, there is a direct argument for **P**(-*H* & *Mo*)=1/2, but **P**(*H*+ & *Mo*) will only be equal to **P**(-*H* & *Mo*) if **P**(*Mo*)=1, which is not correct because awakening on Tuesday is possible. This implies that **P**(*H*+ & *Mo*)≠**P**(-*H* & *Mo*) in general, and that **P**(*H*+ & *Mo*) may be an invalid expression because of the contradiction arising from future-tensed reasoning. An analysis on Tuesday creates even more problems for future-tensed coin tosses (*H*+, *T*+), because on Tuesday the toss will have already occurred, rendering future-tensed expressions nonsense. While Elga's original argument that **P**(*H* & *Mo*) equals 1/3 does not contain any clear errors, the argument entails other consequences that are internally contradictory.

When making the distinction between -*H* and *H*+, the so-called double-halfer argument does not contain an inconsistency, so long as we perform the experiment according to method 1:

P(*H*+)=**P**(*H*+|*Mo*)=**P**(-*H* & *Mo*)=**P**(-*H*)=1/2

The double-halfer can believe that **P**(*H*+|*Mo*)=1/2 without embarrassment — she simply will not use this fact in making an argument for **P**(-*H* & *Mo*).

Titelbaum (2012) has suggested a variation of the Sleeping Beauty problem that creates an interesting dilemma for the double-halfer position. In this variant, a fair coin is also tossed on Tuesday night in addition to the toss on Monday, but unlike the Monday toss, the Tuesday toss does not influence the course of the experiment — the researcher just tosses it on Tuesday for fun. With a fair toss on each day, we can now ask about the subject's credence, upon awakening, that *today's toss* is heads. Titelbaum presents this experiment with the coin being tossed 'at night' at the end of each day. Both tosses are future-tensed.

On the one hand, it seems obvious that your credence in Today's toss being Heads, upon awakening on either day, should be 1/2, and therefore your expectation of Today's toss being Heads will be 1/2 as well: **P**(Today Heads)=**P**(*H*+|*Mo*)**P**(*Mo*)+**P**(*H*+|*Tu*)**P**(*Tu*)=1/2·**P**(*Mo*)+1/2·**P**(*Tu*)=1/2. But the concept of 'Today's toss' in the experiment as suggested by Titelbaum is problematic since it necessitates future-tensed reasoning, and entails that Monday and Heads are both independent and dependent. I contend that we ought reject future-tensed reasoning in this example, and reject the expression **P**(*H*+|*Mo*)**P**(*Mo*) as being invalid. The credence **P**(Today Heads)=1/2 does not derive from formal analysis, but follows from reasoning something like this: whatever day it is, on that day my credence in Heads is 1/2.

A simple adjustment to Titelbaum's variation makes it amenable to past-tensed reasoning. Suppose a fair coin is tossed on Sunday night and on Monday night. For the Sunday night coin, if Heads then you wake once on Monday, and if Tails you wake on Monday and Tuesday. The Monday night coin does not influence the experiment. Now, what is your credence upon awakening that *yesterday's* toss was Heads? It is no longer obvious that **P**(Yesterday Heads)=1/2. This is because the dependencies between Sunday's toss and the day you awaken become palpable.

$$\begin{aligned}\mathbf{P}(\text{Yesterday Heads}) &= \mathbf{P}(Mo|\text{-}H_S)\mathbf{P}(\text{-}H_S)+\mathbf{P}(Tu|\text{-}H_M)\mathbf{P}(\text{-}H_M)\\ &= (1)\cdot(1/2)+\mathbf{P}(Tu)\cdot(1/2)\\ &= 1/2+(1/4)\cdot(1/2)=5/8\end{aligned}$$

$\mathbf{P}(Tu|\text{-}H_M)=\mathbf{P}(Tu)$ because whether you wake on Tuesday is independent of the coin toss the day before by the setup of the experiment, and $\mathbf{P}(Tu)=1/4$ in accordance with classic Halfer reasoning. I will leave it to reader to decide if **P**(Yesterday Heads) ought be equal to **P**(Today Heads).

4 Discussion

The Sleeping Beauty Problem has been truly puzzling given its simplicity. Both the Halfer and Thirder solutions have their appeal and reason. No one has identified a clear error or agreed-upon contradiction in reasoning for either side, although Lewis's claim that we should have a conditional credence of 2/3 for Heads given information that it is Monday, prior to tossing the coin, has been a strong contender for Most Likely to be an Error. However, when we conditionalize on Monday after tossing the coin, this criticism is greatly attenuated. The temporal timing of the coin toss in relation to our evidence seems to play an important role in reasoning about the credences for this problem, and upon taking a closer look at Elga's original argument, it appears to contain a demonstrable contradiction. This contradiction should put the Thirder position on alert, although that is not the only reason for questioning it. I have also suggested that reasoning about repeated runs of the experiment, or multiple copies of the experiment; and then trying to apply those results to the single run case may be invalid. It is entirely possible that a repeated Sleeping Beauty experiment will yield a credence in Heads of 1/3, yet the single run credence remains 1/2 (see Appendix B). I also think the Mirror Argument for 1/2 is compellingly neat and look forward to someone poking holes in it.

The following are some suggestions in this work:

- Reasoning about the total number of awakenings per run differs from reasoning about Heads or Tails, despite the fact that these sets of events are equivalent.

- The two Tails awakenings on Monday and Tuesday are mutually necessary. They are selected together or not at all.

- Selection of awakening events is grounded in the causal selection mechanism of the experiment, where first the side of the coin is selected. The three awakenings are not contemporaneously selectable by a single mechanism.

- Reasoning about the multiple-run frequencies of events does not necessarily apply to the single run case.

- Method 1 and method 2 are not equivalent with respect to probabilistic reasoning, despite the fact that both methods cause the same objective distribution of awakening events.

- Method 2 reasoning leads to a direct contradiction using probabilistic temporal analysis.

- Knowledge that an event has been decided (without knowing the result) can be a type of admissible evidence when updating credences.

It is entirely possible I have led myself astray in trying to resolve this problem.

Appendix A. Biased Sleeping Beauty

Here I show expressions for both the Halfer and Thirder solutions when the coin has an arbitrary chance *p* of landing Heads. I will use the labels *H1*, *T1*, and *T2* from Table 1 to identify each situation. The Halfer solution is straightforward

$\mathbf{P}(H1)=p$

Since $\mathbf{P}(H1)+\mathbf{P}(T1)+\mathbf{P}(T2)=1$, and $\mathbf{P}(T1)=\mathbf{P}(T2)$,

$$\mathbf{P}(T1)=\mathbf{P}(T2)=(1-\mathbf{P}(H1))/2$$
$$=(1-p)/2$$

For the Thirder solution, I will make use of Elga's future-tensed argument, but replace 1/2 by *p*, which gives:

$\mathbf{P}(H1)/(\mathbf{P}(H1)+\mathbf{P}(\mathrm{T1}))=p$

$\mathbf{P}(H1)=p\cdot\mathbf{P}(T1)/(1-p)$

Substituting $\mathbf{P}(T1)=(1-\mathbf{P}(H1))/2$, and solving yields

$\mathbf{P}(H1)=p/(2-p)$, and

$\mathbf{P}(T1)=\mathbf{P}(T2)=(1-p)/(2-p)$

	Halfer	Thirder
HI	p	$p/(2-p)$
T1	$(1-p)/2$	$(1-p)/(2-p)$
T2	$(1-p)/2$	$(1-p)/(2-p)$

Table A1. *Solutions to Biased Beauty using a coin with chance p of landing Heads*

While the intuition that gave us the Halfer solution carriers over directly to the case of arbitrary *p*, the intuition that corresponds to the Thirder solution looses all of its force. There is nothing compelling in the statement of the Sleeping Beauty problem to suggest that $\mathbf{P}(H1)=p/(2-p)$ generally, but a Thirder who accepts Elga's argument ought to support this expression more so than the particular value at $p=1/2$.

Appendix B. Repeated Sleeping Beauty

Suppose that instead of flipping a fair coin just once and waking the subject up one or two times, that the experiment was continued for multiple coin flips. In this variant of the problem, the coin is flipped on day 1, the subject wakes on day 2 or both days 2 and 3 dependent upon Heads or Tails, and kept

asleep on day 4. On day 4, the coin is flipped again, and the subject wakes on day 5 or day 5 and day 6 dependent on the toss, and so forth. The subject's memory is erased after each awakening. Now in this experiment of n tosses, what is the subject's credence **P**(*HEADS*) that he awakens in a Heads awakening? In the following section, I define **P**(*HEADS*)=**P**(It is a HEADS awakening).

The complete set of outcomes is listed in Table B1 for the n=2 case

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
H	A_H		H	A_H	
H	A_H		T	A_T	A_T
T	A_T	A_T	H	A_H	
T	A_T	A_T	T	A_T	A_T

Table B1. Awakening for n=2

Heads-awakenings are designated by A_H and Tails-awakenings by A_T, so **P**(*HEADS*)=**P**(A_H). Instead of flipping a coin on two different days, the researcher might decided to flip the coin two times on Day 1, and simply not do anything on Day 4 (Table B2). It is possibly that this change will influence our reasoning or calculations. Initially I will assume it does not. Each awakening is individually identified.

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
HH	H_1H			HH_1	
HT	H_1T			HT_1	HT_2
TH	T_1H	T_2H		TH_1	
TT	T_1T	T_2T		TT_1	TT_2

Table B2. Awakenings for n=2 individually labeled

B1 Halfer analysis

I propose the following analysis. Before you are put to sleep, each pair of tosses {HH, HT, TH, TT} are equally likely, with credence of 1/4. When you awaken, you are in an experimental run determined by one of those pairs. Suppose the coins landed HH. Then when you awaken, you are either in H_1H or HH_1. A restricted prinicple of indifference (Elga), or a notion of collocation of uncentered possibilities (Lewis), yields that you ought to have equal credence in each possibility, conditional on HH. A similar reasoning applies if the coins landed HT, TH, or TT. Noting that the following equal conditional credences sum to one yields:

(B1) $\mathbf{P}(H_1H|\text{HH})=\mathbf{P}(HH_1|\text{HH})=1/2$

$\mathbf{P}(H_1T|\text{HT})=\mathbf{P}(HT_1|\text{HT})=\mathbf{P}(HT_2|\text{HT})=1/3$

$\mathbf{P}(T_1H|\text{TH})=\mathbf{P}(T_2H|\text{TH})=\mathbf{P}(TH_1|\text{TH})=1/3$

$\mathbf{P}(T_1T|\text{TT})=\mathbf{P}(T_2T|\text{TT})=\mathbf{P}(TT_1|\text{TT})=\mathbf{P}(TT_2|\text{TT})=1/4$

Then by the definition of conditional probability

(B2) $\mathbf{P}(H_1H)=\mathbf{P}(H_1H\ \&\ \text{HH})=\mathbf{P}(H_1H|\text{HH})\mathbf{P}(\text{HH})=(1/2)\cdot(1/4)=1/8$

$\mathbf{P}(H_1T)=\mathbf{P}(H_1T\ \&\ \text{HT})=\mathbf{P}(H_1T|\text{HT})\mathbf{P}(\text{HT})=(1/3)\cdot(1/4)=1/12$

$\mathbf{P}(T_1H)=\mathbf{P}(T_1H\ \&\ \text{TH})=\mathbf{P}(T_1H|\text{TH})\mathbf{P}(\text{TH})=(1/3)\cdot(1/4)=1/12$

$\mathbf{P}(T_1T)=\mathbf{P}(T_1T\ \&\ \text{TT})=\mathbf{P}(T_1T|\text{TT})\mathbf{P}(\text{TT})=(1/4)\cdot(1/4)=1/16$

And so forth, such that

$\mathbf{P}(H_1H)=\mathbf{P}(HH_1)=1/8$

$\mathbf{P}(H_1T)=\mathbf{P}(HT_1)=\mathbf{P}(HT_2)=1/12$

$\mathbf{P}(T_1H)=\mathbf{P}(T_2H)=\mathbf{P}(TH_1)=1/12$

$\mathbf{P}(T_1T)=\mathbf{P}(T_2T)=\mathbf{P}(TT_1)=\mathbf{P}(TT_2)=1/16$

(B3) $\therefore \mathbf{P}(HEADS)=\mathbf{P}(H_1H)+\mathbf{P}(HH_1)+\mathbf{P}(H_1T)+\mathbf{P}(TH_1)=2/8+2/12=5/12$

Although I implemented a halfer-type reasoning to derive this answer, $\mathbf{P}(HEADS)$ does not equal 1/2 when n=2 tosses. The equations above are the 'formal' derivation of the n=2 solution, but consider the following tables to determine a general expression for an arbitrary number of tosses.

Outcome n=2	Total Awakenings	Heads Awakenings
HH	2	2
HT	3	1
TH	3	1
TT	4	0

For n=2 tosses
$\mathbf{P}(HEADS)$
=(2/2+1/3+1/3+0/4)*1/4
=5/12

Outcome n=3	Total Awakenings	Heads Awakenings
HHH	3	3
HHT	4	2
HTH	4	2
THH	4	2
TTH	5	1
THT	5	1
HTT	5	1
TTT	6	0

For n=3 tosses
$\mathbf{P}(HEADS)$
=(3/3+2/4+2/4+2/4+1/5+1/5+1/5+0/6)*1/8
=31/80

We need only add the fractions of Heads-awakenings for each outcome of the coin tosses, and then multiple by $1/2^n$, which is the probability of each outcome for *n* tosses. Noting that each fraction occurs *n* choose *k* Heads times (the respective binomial coefficient) we can deduce that the expression reduces to:

(B4) $$\mathbf{P}(HEADS_n) = \frac{1}{2^n}\sum_{k=0}^{n}\frac{n-k}{n+k}\binom{n}{k}$$

And further point out two interesting values for *n*:

$$\mathbf{P}(HEADS_1) = \frac{1}{2}(1+0) = \frac{1}{2}$$

Which is the standard Halfer solution when *n*=1, where the only possible outcomes are H and T, and

$$\lim_{n\to\infty}\mathbf{P}(HEADS_n) = \frac{1}{3}$$

Which reduces to the standard Thirder solution of 1/3 as *n* approaches infinity, when there are an infinite number of coin tosses. I evaluated this result numerically, although those more able at math may derive the limit directly.

Why is **P**(*HEADS*) less then 1/2 for *n*>1? As *n* gets larger, sequences of coin tosses (outcomes) with approximately equal numbers of Head and Tails tend to dominate the complete set of possible outcomes, in accordance with the binomial coefficients. And within outcomes with approximately equal numbers of Heads and Tails, the ratio of Heads to Tails *awakenings* is approximately 1 to 2, thus the probability of a Heads awakening approaches 1/3 as *n* approaches infinity. Equation (B4) is consistent in that it reproduces the Halfer solution of 1/2 for *n*=1, and further, validates an aspect of Thirder reasoning: in the long run of a repeated experiment, your credence in Heads ought to approach 1/3. It also demonstrates that sometimes the result of a repeated experiment or long-run analysis does not apply to the single run case.

B2 Thirder inspired analysis

So much for a Lewisian inspired Halfer solution to repeated Sleeping Beauty. What happens when we try to apply Elga's reasoning to this problem? The restricted principle of indifference that gave us (B1) carries over to this analysis, yet instead of using the equality of conditional credences, I will use the unconditional equalities pace Elga and Lewis:

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
HH	H_1H			HH_1	
HT	H_1T			HT_1	HT_2
TH	T_1H	T_2H		TH_1	
TT	T_1T	T_2T		TT_1	TT_2

(T1) $\mathbf{P}(H_1H)=\mathbf{P}(HH_1)$

$\mathbf{P}(H_1T)=\mathbf{P}(HT_1)=\mathbf{P}(HT_2)$

$\mathbf{P}(T_1H)=\mathbf{P}(T_2H)=\mathbf{P}(TH_1)$

$\mathbf{P}(T_1T)=\mathbf{P}(T_2T)=\mathbf{P}(TT_1)=\mathbf{P}(TT_2)$

Since the credences of these awakenings must sum to one:

(T2) $2\cdot\mathbf{P}(H_1H)+3\cdot\mathbf{P}(H_1T)+3\cdot\mathbf{P}(T_1H)+4\cdot\mathbf{P}(T_1T)=1$

So far, in (T2) there are four unknowns and only one equation, so the solution is underdetermined. Elga used the following reasoning to derive additional equations to solve the problem for n=1:

> Your credence that you are in H1 would then be your credence that a fair coin, soon to be tossed, will land Heads. It is irrelevant that you will be awakened on the following day if and only if the coin lands Tails — in this circumstance, your credence that the coin will land Heads ought to be 1/2. But your credence that the coin will land Heads (after learning that it is Monday) ought to be the same as the conditional credence P(H1|H1orT1). So P(H1|H1orT1)=1/2…

We can first apply a similar reasoning for the first toss in a sequence of two tosses. Suppose you know it is Day 1 and that the first coin was not yet tossed. Then your credence that a fair coin, soon to be tossed, will land Heads is 1/2. This ought to be the same as the conditional credence $\mathbf{P}(H_1H$ or $H_1T|H_1H$ or H_1T or T_1H or $T_1T)$. So

(T3) $\mathbf{P}(H_1H$ or $H_1T|H_1H$ or H_1T or T_1H or $T_1T)=1/2$

$2\cdot\mathbf{P}(H_1H)+2\cdot\mathbf{P}(H_1T)=\mathbf{P}(H_1H)+\mathbf{P}(H_1T)+\mathbf{P}(T_1H)+\mathbf{P}(T_1T)$

$\mathbf{P}(H_1H)+\mathbf{P}(H_1T)-\mathbf{P}(T_1H)-\mathbf{P}(T_1T)=0$

Now there are two equations and four unknowns, which is still undetermined. A similar reasoning can be applied to the second toss, assuming that you know the second toss is fair and has not yet occurred

(T4) $\mathbf{P}(HH_1 \text{ or } TH_1 | HH_1 \text{ or } TH_1 \text{ or } HT_1 \text{ or } TT_1)=1/2$

$\mathbf{P}(HH_1)+\mathbf{P}(TH_1)-\mathbf{P}(HT_1)-\mathbf{P}(TT_1)=0$

Recalling (T1) and substituting

$\mathbf{P}(H_1H)+\mathbf{P}(T_1H)-\mathbf{P}(H_1T)-\mathbf{P}(T_1T)=0$

Taken together, we have a system of three equations and four unknowns. This system is underdetermined, and either has zero or an infinite number of solutions. Where might an additional constraint be found? Elga's argument perhaps can be extended to the case of two coin tosses, where your credence in getting two Heads for two coins yet to be tossed is 1/4, which ought to be equivalent to

(T5) $\mathbf{P}(H_1H | H_1H \text{ or } H_1T \text{ or } T_1H \text{ or } T_1T)=1/4$

$4\cdot\mathbf{P}(H_1H)-\mathbf{P}(H_1H)-\mathbf{P}(H_1T)-\mathbf{P}(T_1H)-\mathbf{P}(T_1T)=0$

$3\cdot\mathbf{P}(H_1H)-\mathbf{P}(H_1T)-\mathbf{P}(T_1H)-\mathbf{P}(T_1T)=0$

Giving us another equation, enough to solve the problem. The next three equations follow as well, from similar expressions:

$3\cdot\mathbf{P}(H_1T)-\mathbf{P}(H_1H)-\mathbf{P}(T_1H)-\mathbf{P}(T_1T)=0$

$3\cdot\mathbf{P}(T_1H)-\mathbf{P}(H_1H)-\mathbf{P}(H_1T)-\mathbf{P}(T_1T)=0$

$3\cdot\mathbf{P}(T_1T)-\mathbf{P}(H_1H)-\mathbf{P}(H_1T)-\mathbf{P}(T_1H)=0$

Now the system is overdetermined with seven equations and four unknowns. Another four equations can be derived by applying the reasoning in (T5) to the awakenings on Day 5, but those are identical to those in (T5). The system of equations can be written in matrix form:

(T6)
$$\begin{pmatrix} 2 & 3 & 3 & 4 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} H_1H \\ H_1T \\ T_1H \\ T_1T \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Performing some linear algebra, we will find that the rank of the first matrix is 4, and that we can select a set of four linearly independent rows, and solve for the four unknowns, giving

$$\mathbf{P}(H_1H)=\mathbf{P}(H_1T)=\mathbf{P}(T_1H)=\mathbf{P}(T_1T)=1/12$$

Adding together the four Heads-awakenings yields

$$\mathbf{P}(HEADS_2)=\frac{1}{3}$$

I will not attempt to formally generalize this result to arbitrary *n*, but it would be easy enough to apply a similar analysis to any size *n*, generate the matrix and solve. I conjecture that for arbitrary *n*, a generalized Thirder argument for the Repeated Sleeping Beauty experiment would give

$$\mathbf{P}(HEADS_n)=\frac{1}{3}$$

I don't suspect this would surprise anyone. The path to the solution is not particularly elegant and comes off as a bit contrived, but that could be my fault.

Acknowledgements

I thank my wife Jean for not getting upset that I spent so much time thinking about a problem that seems quite silly to her.

Marc Burock
burocksmail@gmail.com

References

Bostrom, Nick. [2007]: "Sleeping beauty and self-location: A hybrid model". *Synthese*, 157(1), 59–78.

Elga, Adam. [2000]: "Self-locating belief and the Sleeping Beauty problem". *Analysis* 60: 143–47.

Kierland, Brian, and Bradley Monton. [2005]: "Minimizing Inaccuracy for Self-Locating Beliefs." *Philosophy and Phenomenological Researc*h, vol. 70, no. 2, pp. 384–395.

Lewis, David. [1980]: "A subjectivist's guide to objective change." In Richard C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, Volume II. Berkeley: University of California Press. pp. 263-293.

Lewis, David. [2001] "Sleeping Beauty: Reply to Elga". *Analysis*, 61(3), 171–176.

Liao, Shen-Yi. [2012]: "What Are Centered Worlds?" *The Philosophical Quarterly*, vol. 62, no. 247, pp. 294–316.

Titelbaum, Michael G. [2012]: "An Embarrassment for Double-Halfers." *Thought: A Journal of Philosophy*, vol. 1, no. 2, pp. 146–151.

Titelbaum, Michael G. [2013]: "Ten Reasons to Care About the Sleeping Beauty Problem." *Philosophy Compass*, vol. 8, no. 11, pp. 1003–1017.