Synthese
https://doi.org/10.1007/s11229-017-1647-x

The Snow White problem

Sylvia Wenmackers[1]

Received: 14 June 2017 / Accepted: 25 November 2017

Abstract The Snow White problem is introduced to demonstrate how learning something of which one could not have learnt the opposite (due to observer selection bias) can change an agent's probability assignment. This helps us to analyse the Sleeping Beauty problem, which is deconstructed as a combinatorial engine and a subjective wrapper. The combinatorial engine of the problem is analogous to Bertrand's boxes paradox and can be solved with standard probability theory. The subjective wrapper is clarified using the Snow White problem. Sample spaces for all three problems are presented. The conclusion is that subjectivity plays no irreducible role in solving the Sleeping Beauty problem and that no reference to centered worlds is required to provide the answer.

Keywords Probability · *de se* beliefs · Observer selection effects · Bayesianism · Sleeping Beauty problem

Two paradoxes are better than one; they may even suggest a solution.
– Edward Teller (Teller et al. 2002, p. 135)

1 The Snow White problem

Let me present a new problem, inspired by the Sleeping Beauty problem (Elga 2000) and intended to clarify the latter.

✉ Sylvia Wenmackers
sylvia.wenmackers@kuleuven.be

[1] Institute of Philosophy, Centre for Logic and Analytic Philosophy, KU Leuven, Kardinaal Mercierplein 2, Bus 3200, 3000 Leuven, Belgium

Published online: 04 December 2017

THE SNOW WHITE PROBLEM
Snow White (SW) is about to meet the witch (W), who will offer her an apple. Before they meet, W will toss a fair coin: if it lands heads, she will offer SW a poisoned apple, which will kill her instantly; if the coin lands tails, W will offer SW an enchanted apple, which will make everyone live happily ever after. The huntsman has warned SW of W's plan, but she decides that she will accept and eat the apple anyway, putting her life at risk in the hope of an enchantment.
Now assume that SW has just finished eating the apple offered to her by W. She finds herself to be alive. Which degree of belief should she now rationally assign to the possibility that the coin toss has produced heads?

The answer is zero. SW is certain that the coin has not landed heads, for else, she would be dead by now.[1]

It may seem as though there is also a case to be made for the answer 1/2. Indeed, there is a fair coin involved, which has an equal prior probability of landing heads:

$$P(H) = P(\neg H) = \frac{1}{2},$$

where H refers to the event that the coin lands heads and ¬H to the event that the coin lands tails.[2]

On the other hand, there is additional information in the story, which requires us to consider a conditional probability. The crucial information, which will function as the conditioning event, is that SW has survived eating the apple. So, the answer to the puzzle can be found via this conditional probability:

$$P(H \mid A \cap L) = \frac{P(H \cap A \cap L)}{P(A \cap L)} = \frac{P(\emptyset)}{P(A \cap L)} = 0,$$

where A refers to the event that SW receives and eats W's apple and L refers to the event that SW lives.

Since SW learns that she has survived eating the apple (and nothing more), the above conditioning event is in agreement with her current information. Therefore, it is rational for her to align her degree of belief (or 'credence') with this conditional probability.

Of course, SW may describe the conditioning event as "I am alive", something which rarely counts as new information that is relevant to update one's degrees of belief upon. However, the whole set-up of the story clearly requires us to do so for this problem.

Although it may appear otherwise, solving this puzzle does not teach us anything essential about the interplay between (subjective) probability and *de se* beliefs. After

[1] This problem is so trivial that it barely deserves the name. It involves a biconditional and elementary logic suffices to find the answer. Yet, I will apply probability theory to it to clarify the subjective aspect of the Sleeping Beauty problem.

[2] Events will be represented as sets throughout.

all, it does not matter who determines the conditional probability, as long as the modelling agent has exactly the same information as SW has at the time of interest. SW can do this herself, but also the huntsman, who has not witnessed the coin toss, can deduce that the coin must have landed tails (with P = 1) when he observes SW eating the apple and surviving it.

Moreover, SW can calculate the above conditional probability at any time, but since the conditioning event does not apply to her before she has eaten the apple, it is not rational to align her degrees of belief with it prior to meeting W. All the relevant information she possesses before eating the apple is that W uses a fair coin. So, at that point in time it is rational for SW to keep her degree of belief aligned with the prior probability, 1/2.

Unlike SW, the huntsman may also observe SW's death after eating the apple, in which case he can deduce that the coin must have landed heads. The relevant conditional probability is:

$$\mathrm{P}(\mathrm{H} \mid \mathrm{A} \cap \neg \mathrm{L}) = \frac{\mathrm{P}(\mathrm{H} \cap \mathrm{A} \cap \neg \mathrm{L})}{\mathrm{P}(\mathrm{A} \cap \neg \mathrm{L})} = \frac{\mathrm{P}(\neg \mathrm{L})}{\mathrm{P}(\neg \mathrm{L})} = 1.$$

Before the coin toss, SW can compute this conditional probability, too. The main difference between her and the huntsman is that SW can never be in a situation such that she has to align her degrees of belief with this conditional event. In this limited sense, the problem exhibits a form of observation selection bias. However, this does not affect the degree of belief that the puzzle asks for.

Still, we may suspect that a contradiction is lurking right beneath the surface. Given that SW knows beforehand that, once given the opportunity, she will lower her degree of belief in heads after eating the apple from 1/2 to 0, shouldn't she do so already? If the answer is 'yes', this contradicts the assumption of a fair coin. Should we now argue that SW's degree of belief ought to stay 1/2 after all, in order to avoid this contradiction? The answer is 'no' to both questions: SW should not revise her degrees of belief beforehand, for she has not been given any evidence yet. If she is in a position to revise her degrees of belief afterwards, then she learns that the apple was not poisonous, which only occurs if the coin landed tails. There is no contradiction, not even an implicit one.

Weisberg (2005) considers a similar example of a firing squad, due to Leslie, in which the prior probability of survival is much smaller than 1/2. He argues that the fact that one cannot observe one's own non-survival does not constitute a full-blown observation selection effect, but merely a conditional: "If I observe whether I survive, I will observe that I survive" (p. 816). This applies to the Snow White problem, too: *if* SW survives, then she observes *that* she survives. Weisberg also observes the essential agreement between a first- and third-person perspective in this case. He further brings this to bear on the weak anthropic principle and the issue of cosmological fine-tuning, which I will not go into here.

At this point, I hope that the reader has been convinced that learning something of which one could not have learnt the opposite—not because it could not have been otherwise, but because of the aforementioned (conditional) selection bias—can lower

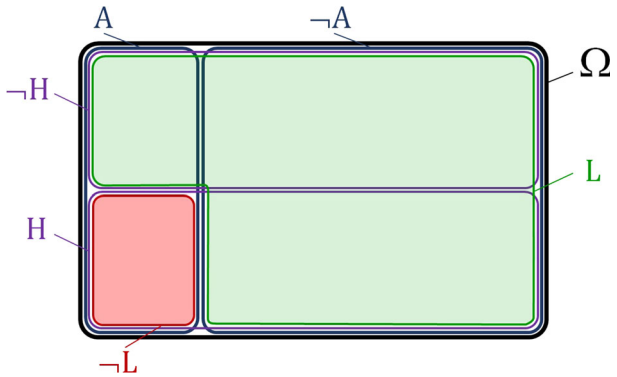

Fig. 1 Venn diagram representing a sample space (Ω) for analysing the Snow White problem, with three partitions of it: {A, ¬A}, {H, ¬H}, and {L, ¬L}. The areas of intersections suggest probability measures of various events. It can immediately be seen that $P(\neg H \mid A \bigcap L) = 1$

one's posterior probability assignment to a fair coin having landed heads (even to zero).

1.1 Sample space for the Snow White problem

Probabilities are assigned to events, which are subsets of a sample space. It is straightforward to present a sample space for the Snow White problem. The symbols for events, which have already been introduced, correspond to members of partitions of this sample space. While the events can be described using tensed statements, indexicals, and demonstratives by specific agents in particular contexts, referring to them by the symbols is always possible.
As can be seen in Fig. 1, the sample space (Ω) can be partitioned in at least three ways:

(1) {A, ¬A}, representing whether or not SW receives and eats W's apple,
(2) {H, ¬H}, representing whether or not the coin toss lands heads, and
(3) {L, ¬L}, representing whether or not SW is alive.

Most of the time, SW does not receive any potentially poisoned apple from W (¬A; righthand-side of Ω in Fig. 1), and her being alive is independent from coin toss results. (Since SW is fictional, we may assume her probability of dying to be zero when she receives no apple from W.)

In this simple case, the event ¬L ("SW is dead") is equivalent to $H \bigcap A$ ("The coin lands heads & SW receives and eats W's apple"), but in general, sets from various partitions may intersect in different ways. In addition, despite the two-dimensional representation, the sample space is best thought of as a high-dimensional set, which may include spacetime coordinates as variables.

1.2 Analogy between SW and SB

For self-containedness, I now give a rephrasing of the Sleeping Beauty problem.

> Sleeping Beauty and a wicked fairy
> Sleeping Beauty (SB) is about to be put under a sleeping spell by a wicked fairy (WF). SB will sleep for a long time during which WF will wake her up either for one or for two brief periods. As soon as SB sleeps, WF will toss a fair coin: if it lands heads, WF will wake SB once; if it lands tails, WF will wake SB twice. After each waking period, WF administers a poison to SB, such that SB will not remember that she has been awake. SB is informed of all these things before she is put to sleep.
> Now assume that SB has just been awoken by WF. SB does not receive any clues from which she could deduce whether this is the first or the second awakening. Which degree of belief should she now rationally assign to the possibility that the coin toss has produced heads?

Let us start by elucidating two analogies between this scenario and the SW problem. First, the fact that one is conscious (SW still alive; SB currently awake) rarely counts as a relevant piece of information when evaluating probabilities. Nevertheless, the contrived set-up of both stories requires us to do so for these puzzles. Second, the fact that one cannot observe oneself to be unconscious (SW dead; SB asleep) is not essential for calculating the relevant probabilities, nor is the fact that the subjects in these puzzles have to evaluate information pertaining to themselves. To illustrate this, we may introduce an observer, who has the same information as the subjects: since we introduced the huntsman for the Snow White problem, let us consider the prince for the Sleeping Beauty problem. Suppose that the prince receives a picture of SB taken at a random time during the sleeping spell: if he happens to see SB awake, he may lower his probability of the coin having landed heads to 1/3—as should SB herself when she finds herself to be awake during the procedure. Of course, it is far more likely that the prince will see a picture in which SB is asleep, in which case he can raise his degree of belief in the coin having landed heads slightly above 1/2. (The exact amount depends on how long a waking episode lasts as compared to the total duration of the spell.) This shows that the first-person aspect in the Sleeping Beauty is 'transparent', in the sense that the relevant events can be described by third-person propositions, without altering the relevant conditional probability. The set-up of WF's spell is such that SB is maximally uncertain about what time it is when she finds herself awake during the spell. Roughly, the evidence that she thereby receives—and that only SB can gloss as "I am awake now"—is equivalent to: "at a randomly sampled moment during WF's spell, SB is awake". However, this mixes an event and its probability; see Sect. 4 for a more precise treatment that disentangles these aspects.

There is, of course, an important dissimilarity between the Sleeping Beauty problem and the Snow White problem: whereas SB can take it for granted that she will wake up at least once, SW cannot be sure that she will survive. This makes the case for 1/2 weaker in the Sleeping Beauty problem than in the Snow White problem. To overcome this disanalogy, I now discuss the combinatoric engine of the Sleeping Beauty problem separately.

2 Bertrand's boxes paradox

Let us first verify that the answer to the Sleeping Beauty problem is indeed 1/3, as originally claimed (Elga 2000). To this aim, I invoke an analogy—with an adapted form of Bertrand's boxes paradox (Bertrand 1889).[3] This analogy has also been observed by (Rosenthal 2009), although he admits that it needs "some sort of reformulation or philosophical analysis" (p. 36), before the relevant conditional probabilities can be brought to bear on the SB case. Such an analysis is precisely what I aim to offer here.

Although Bertrand presented his boxes puzzle as a paradox of probability,[4] contemporary mathematicians view the problem as one with a unique and indisputable answer.[5]

> VARIATION ON BERTRAND'S BOXES PROBLEM
> Bertrand (B) is about to meet Joseph (J), who will carry a box that contains one thousand medals. Before they meet, J will toss a fair coin: if it lands heads, J will take a box that contains 1 gold medal and 999 silver medals; if it lands tails, J will take a box that contains 2 gold medals and 998 silver ones. When they meet, B has to pick a medal from the box without looking into it.
> Now assume that B has just picked a medal from the box that J carried and it turns out to be a gold medal. What is the probability that the coin has produced heads?

The answer is 1/3.

A correct way of reasoning is as follows. Initially, each of the 2000 medals has the same probability of being selected (this equiprobability is due to the combination of J's use of a fair coin and B's selection of a medal without looking). We are asked to assume that the first medal is a golden one, which occurs in three out of these 2000 cases. Of these three ways to select a gold medal (which are all equally likely), the medal comes from the box containing just one gold medal in one out of three cases. So, the probability that the selected medal comes from the box containing a gold medal and 999 silver ones is 1/3. J carries this box only if the coin landed heads. Hence, the probability that the coin landed heads is also 1/3.[6]

It may seem as though there is also a case to made for the answer 1/2: after all, J uses a fair coin and both boxes contain at least one gold medal. However,

[3] Bertrand's paradoxes are well-known in the contemporary literature as illustrations of the difficulty in applying the principle of indifference. Bertrand's boxes problem is a particularly simple example, appearing early in the first chapter—even before the famous chord problem—involving three boxes each containing two medals (either both silver, both gold, or one of each). It is not to be confused with the 'boxes factory', which is a more recent paradox (also related to indifference) due to van Fraassen.

[4] A similar problem, which asks for the probability of obtaining at least one heads in two consecutive tosses with a fair coin, even got the famous mathematician d'Alembert fooled (see, e.g., Henry 2005).

[5] Illustrative is the fact that the boxes problem served as a template for a question in the Dutch national science quiz of 2011, which is organized by the Netherlands Scientific Research Organization as a popularization effort for mathematics and science (NWQ 2011).

[6] Observe that the solution does not depend on the fact that the boxes each contain a thousand medals; what does matter is that both boxes contain an *equal* number of at least two medals, exactly one versus two of which are gold.

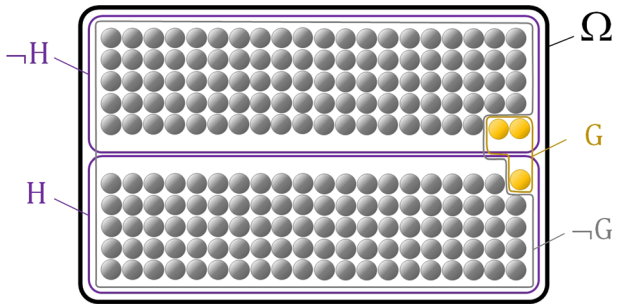

Fig. 2 Venn diagram representing a sample space (Ω) for analysing the variant of Bertrand's boxes problem, with 200 medals rather than 2000, with two partitions of it: {H, ¬H} and {G, ¬G}. The circles indicate atomic events, which are all equiprobable. It can immediately be seen that $P(H \mid G) = 1/3$

the probability that the medal comes from the box containing two gold medals is twice as large as the probability that it comes from the box containing only one gold medal.[7]

2.1 Sample space for Bertrand's boxes problem

The above claims can be checked with the sample space given in Fig. 2, albeit for 200 medals rather than 2000. H and ¬H stand for the events that the coin lands heads or tails, respectively. G and ¬G stand for the events that a gold or silver medal are selected, respectively.

2.2 Analogy between B and SB

Let us now explore the analogy between the boxes problem and the Sleeping Beauty problem. Obviously, the gold medals are intended to correspond to "SB is awake" and the silver medals to "SB is asleep". The fact that each box contains a gold medal does not alter the fact that B's drawing one ought to change the degree of belief that we assign to the coin having landed heads. Likewise, the fact that SB knows that she will be awake at least once during the sleeping spell, does not alter the fact that finding herself to be awake ought to change her degree of beliefs.[8]

[7] Observe that if one box contains two medals and the other just one (and nothing else), and all three are gold, nothing can be learnt from picking out a single gold medal: the conditional probability of the coin having landed heads remains equal to the unconditional value, 1/2. This situation is not structurally similar to the original one—although thinking so is a common mistake (*cf.* footnote 4). Crucially, also the Sleeping Beauty problem does *not* have this structure, as will be demonstrated below.

[8] If you are bothered by the fact that SB is guaranteed to be awake at least once, whereas drawing a gold medal is not at all, see the Wicked Joseph scenario below.

To make the analogy more explicit, we can model the Sleeping Beauty problem as follows. Let us assume that the duration of one waking period lasts 1/1000 of the total duration of the spell.[9] Then the probabilities in the Sleeping Beauty problem are:[10]

$$\begin{aligned} \mathrm{P}(\mathrm{H}) &= \mathrm{P}(\neg\mathrm{H}) = \frac{1}{2}, \\ \mathrm{P}(\mathrm{W}) &= \frac{3}{2000}, \\ \mathrm{P}(\neg\mathrm{W}) &= \frac{1997}{2000}, \end{aligned}$$

$$\mathrm{P}(\mathrm{H} \mid \mathrm{W}) = \frac{\mathrm{P}(\mathrm{H} \cap \mathrm{W})}{\mathrm{P}(\mathrm{W})} = \frac{1}{3}.$$

where W stands for SB is awake and ¬W for SB is asleep.[11]

Since SB learns that she is awake during the spell (and nothing more), it is rational for her to align her degree of belief in the coin having landed heads with the conditional probability P(H|W), which does not depend on any assumption concerning the duration of the waking periods. So, 1/3 is the answer to the puzzle.

Although it is not required for solving the puzzle, it is helpful for the further discussion also to compute the following conditional probability:

$$\mathrm{P}(\mathrm{H} \mid \neg\mathrm{W}) = \frac{\mathrm{P}(\mathrm{H} \cap \neg\mathrm{W})}{\mathrm{P}(\neg\mathrm{W})} = \frac{999}{1997} > \frac{1}{2}.$$

2.3 Lucid beauty

It might be objected that the above reconstruction of the puzzle, in which SB is awake a specified (albeit arbitrary) fraction of time, is crucially different from the original puzzle. After all, the awakenings need not be equally long. It is true that the model introduces some superfluous structure, but it is only there to simplify the demonstration; the solution does not depend on it.[12]

[9] The value of the denominator is set to 1000 for definiteness only; it will become clear that nothing hinges on this choice. Any denominator larger than or equal to two is fine (to allow for two awakenings in the duration of the spell). This independence is similar as before (*cf.* footnote 6). The crucial part is that the denominators are equal (*cf.* footnote 9), which is motivated by the fact that the spell's total duration is independent of the coin toss.

[10] Of course, SB may describe the conditioning event as "I am awake". This makes no essential difference, as was argued in Sect. 1.2 on the Snow White problem.

[11] The glosses "SB is awake" for W and "SB is asleep" for ¬W are shorthand and are supposed to include background information such as "SB is under WF's spell" (Sect. 4.1), which, crucially, also encodes how the moment is sampled (Sect. 4.2).

[12] A more complicated version takes into account all possible durations and a higher-order probability function over them. Yet, the overall solution averages out to the one presented in the 'symmetric' case (in which all awakenings have equal duration).

Moreover, even if you reject the previous model, the main point can be demonstrated without it. Suppose that SB was a lucid dreamer, occasionally able to observe herself asleep (without memory of earlier lucid or waking moments during the spell): during such a lucid episode, she would have to assign a probability of more than 1/2 to the coin having landed heads. This probability can be has high as 1 (if SB is remains awake for half of the spell upon each awakening), but if SB is only woken up briefly (as stipulated by Elga 2000) it will be only slightly more than 1/2 (as is the case in the numerical example above). Now the crucial observation: even if SB was a lucid dreamer, she would have to assign a probability of strictly less than 1/2 to the coin having landed heads when she found herself to be awake. (It does require a numerical model to demonstrate that it is exactly 1/3.) In the original version, where Sleeping Beauty does not have the ability of lucid dreaming, finding herself *not asleep* is like hearing the dog that didn't bark in the night: it is hard to notice, but once learnt, it solves the puzzle.

2.4 Wicked Joseph

I already argued that the fact that SB cannot witness herself asleep does not alter the fact that she receives new evidence from being awake. This point can also be illustrated by tinkering further with the boxes problem. Suppose that J is an evil wizard, whose boxes contain a poisonous gas, which renders bystanders unconscious and erases their memory of the previous minutes. If B happens to pick a silver medal, the gas is released immediately; if he picks a gold medal, the release of the gas is delayed. Only when he picks a gold medal, he will be able to realize (briefly) which metal he has picked; in all cases, he forgets it afterwards. Also, assume that J lets B continue picking medals until the box is empty and that B cannot tell whether it is the first, the second, … or the thousandth medal he is about to pick. If B knows in advance that there will be thousand drawings and that he will only be conscious after having drawn a gold medal, then he also knows that he will experience drawing one gold medal at least once in any case.

None of these alterations to the boxes problem influence the relevant conditional probability. So, B's degree of belief that the coin landed heads still ought to drop from 1/2 to 1/3 as soon as he does pick a gold medal. Likewise, SB's degree of belief should change as soon as she wakes up during the procedure.[13]

For the Snow White problem, it was discussed why SW should not update before eating W's apple. While SW does not know that she will be able to receive any evidence, both SB and B do know in advance that they will receive evidence at least once and what the content of this first piece of evidence will be. For B, this piece of evidence can be described as "a randomly sampled coin in the wicked Joseph set-up is golden". Still, the learning only occurs when they find themselves conscious during the procedure. If SB is in a position to revise, then she learns that she is awake at a randomly sampled moment during WF's spell, which is twice as probable on heads than on tails. Similarly for B. (See Sects. 3.2 and 4 for further discussion of these points.)

[13] Since I argue that something is learnt in both cases, van Fraassen's reflection principle (1984) does not apply to either scenario.

3 Indexicals and demonstratives versus intra-model labels

At this point, let me make explicit the purpose of the previous two sections. There have been two main types of responses to the Sleeping Beauty problem:[14] type (1) responses state that SB learns nothing new by waking up and hence it is irrational to update her degrees of belief; type (2) responses state that we need to add something (namely, centered worlds and update rules for them, as explained below) to the Bayesian formalism to capture the propositional content of self-locating beliefs. The arguments in my paper are intended to reply to both.

Replies to (1) are not new. Weintraub (2004) was the first to argue that SB does learn new evidence upon awaking during the spell, namely that she is awake now. I think this reply was on the right track, but I want to make crystal-clear how this can be relevant evidence. In particular, I presented the Snow White problem to answer to (1) and Bertrand's boxes problem to answer to (2); the Lucid Beauty and Wicked Joseph variants add to my reply to (1).

In this section, I will elaborate on my reply to (2) and add to Weintraub's reply to (1) that the evidence need not be in indexical or demonstrative form—thereby connecting the replies to (1) and (2).

3.1 Centered worlds

There may be readers who find my approach so obvious,[15] that they wonder why anyone would reject it. So, let me first try to explain why some philosophers indeed reject it,[16] before replying to them.

The Sleeping Beauty problem was intended as a puzzle about self-locating beliefs, anthropic reasoning, and subjective probabilities: this is clear from the context in which Zuboff started thinking about it (Zuboff 1990; as recounted in: Zuboff 2009). Initiated by the publication of Elga (2000), the problem has generated a large and interesting body of literature that continues to focus on these aspects. Philosophers have many reasons to care about the Sleeping Beauty problem (see, e.g., Titelbaum 2013 for an overview). From a mathematical point of view, however, the mechanism for *solving* the Sleeping Beauty problem does not require any additional rules for updating degrees of self-locating beliefs. It will be shown that there is a lossless translation from first-person to third-person statements and that an outsider with identical information will arrive at the same conclusions.

The combinatorial engine I presented may be regarded as crucially different from the original Sleeping Beauty puzzle, because it replaces uncertainty about centered worlds by uncertainty about which possible world obtains. See for instance (Meacham 2008, p. 249) for a brief description of the distinction between *de dicto* and *de se* propositions, originally due to Lewis (1979). Whereas *de dicto* propositions describe what the world is like (modelled by 'possible worlds'), *de se* propositions describe

[14] For a recent, brief overview of the vast literature on the Sleeping Beauty problem, see Winkler (2017).

[15] For instance, a mathematician like Rosenthal (2009) might.

[16] For instance, I suspect Meacham (2008) and Titelbaum (2013, 2014) would reject it.

when or where the subject is in that world (modelled by 'centered worlds'). When the indexicals in *de se* propositions can be replaced by *de dicto* propositions, they are said to be reducible. But many authors think that the replacement is not possible in cases that involve agents that do not know who they are in the world or at which location or at what time they find themselves: such scenarios seem to involve essential uncertainty over centered worlds. In that case, the scenario is said to involve irreducibly *de se* propositions, and when the probabilities about such propositions change that is called an irreducibly *de se* belief change. Some of these scenarios involve passing of time and changes of truth values of propositions over time. Moreover, many of these scenarios involve forgetting, which cannot be modelled by Bayesian updating (that only takes into account receiving additional information).

Elga (2000) and nearly all commentators adopted Lewis's (1979) distinction and agree that the puzzle involves irreducibly *de se* belief change. It then remains to be discussed how such degrees of belief ought to be revised, since probability theory is silent on this. This is precisely the reason why the Sleeping Beauty problem has received so much attention in the literature: as a test case for demonstrating the vices and virtues of proposed revisions.

I will follow Titelbaum's (2014; Chapter 10) exposition of an extension of the Bayesian framework intended to deal with updating on evidence that can contain context-dependent, indexical information. Titelbaum calls his reconstruction the 'HTM approach', named after Halpern (2005), Tuttle, and Meacham (2008); Titelbaum himself rejects it. On the HTM approach, updating degrees of belief becomes a two-step procedure: new evidence is first used to redistribute one's degrees of belief over centered worlds, then over uncentered worlds. It is a hierarchical procedure in the sense that the second step does not change the degrees of belief over centered worlds that resulted from the first step. In addition, the HTM approach agrees with a "relevance-limiting thesis" (Titelbaum 2014, p. 232): changes in centered evidence (that only eliminate centered possible worlds) should not influence degrees of belief assigned to uncentered possible worlds. Titelbaum explains that this proposed constraint only applies to scenarios in which agents track their evolving spatio-temporal location in otherwise entirely predictable situations. He also conjectures that this principle captures a central intuition motivating many other 'halfers' about the SB problem.

However, I reject that the Sleeping Beauty problem requires us to go beyond the usual Bayesian approach to probability.[17] I already gave the positive part of my argument above, demonstrating that the *de se* belief change in the Sleeping Beauty problem is reducible by considering a third-person perspective or a lucid dreamer, allowing SB to apply probability theory and Bayes' rule to her situation directly.[18]

[17] To be clear, there may be other ways to solve the Sleeping Beauty problem and it may be interesting to consider the subjectivity in the Sleeping Beauty problem, but doing so is not necessary to solve it. Other problems that involve subjective or temporal beliefs may need additional principles for belief change, such as Bradley's (2011b) mutation principle. The Sleeping Beauty problem, however, can be solved by the usual Bayesian formalism alone.

[18] Though I argued along different lines, the conclusion agrees with that of Bradley (2011a): learning where we are in the world gives rise to essentially similar observation selection effects as learning what the world is like by sampling from a population.

My main point is that the information of centered worlds should not be supplemented to the sample space, but should be read off from it using, possibly incomplete, knowledge about who, where, and when a particular agent can possibly be. It is never necessary to conditionalize on statements with indexicals: the sample space requires restricting on the corresponding possible identities, times, and locations instead. This can be done equally well whether starting from a first- or third-person perspective (although the information each modelling agent has can differ, of course). If additional information from centered worlds seems necessary to represent dynamics of *de se* beliefs, the sample space was simply not rich enough to begin with.

I believe that the introduction of centered worlds (as something not reducible to the sample space) is plausibly due to a misunderstanding of the term 'possible worlds' (and 'sample space') as used in probability theory: it means (the set of) all possibilities relevant for the problem at hand, not just possible states of an external, observer-independent world. (See also footnote 19.)

3.2 Main diagnosis of the confusion

Let me now give my diagnosis of the main source of confusion stirred up by the Sleeping Beauty problem and similar puzzles. At any point in time, possible pieces of evidence can be considered hypothetically and the relevant conditional probabilities can be computed with Bayes' theorem. It is only when the relevant agent actually receives a piece of evidence, that the agent is supposed to update his or her degrees of belief by applying Bayes' rule. In this limited sense, receiving evidence and being in a position to update according to Bayes' rule has an irreducibly indexical aspect to it. By indexing probability functions, the Bayesian formalism allows us to represent different agents receiving different pieces of evidence and each agent accumulating evidence over time. It does not follow that the evidence itself necessarily has an indexical aspect to it.

In the Sleeping Beauty scenario and related puzzles, these two aspects become mixed: the agent that is supposed to be doing the Bayesian modelling (SB) is also the subject of the probabilistic scenario she is being asked to model. It has widely been assumed that the evidence SB receives is irreducibly *de se* (which is not captured by the Bayesian formalism). However, my analysis demonstrates that the only indexical aspect in the Sleeping Beauty scenario is that of the updating itself (ever present in Bayesianism), which does not require any extension of the formalism. Until convincingly proven otherwise, we can continue to assume that the Bayesian formalism is not in need of amendment with update rules for centered evidence.

4 What has been forgotten so far

In this section, first I comment on a curious omission in the literature on the Sleeping Beauty problem. Then I briefly comment on the role of forgetting in the problem.

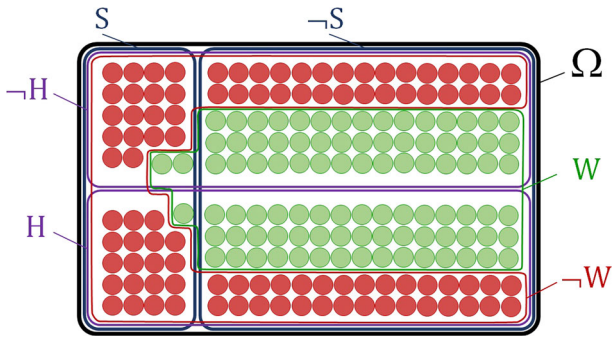

Fig. 3 Venn diagram representing a sample space (Ω) for analysing the Sleeping Beauty problem, with three partitions of it. The areas of intersections suggest probability measures of various events. It can immediately be seen that $P(H|W \bigcap S) = 1/3$

4.1 Sample space for the Sleeping Beauty problem

What is curiously absent from standard presentations and purported solutions of the Sleeping Beauty puzzle is a sample space. I present one in Fig. 3. Similar to my analysis of the Snow White problem, this sample space (Ω) can be partitioned in at least three ways: (1) whether or not SB is under a spell by the WF, (2) whether or not the coin toss landed heads, and (3) whether or not SB is asleep. Most of the time, SW is not undergoing a spell by WF (¬S; righthand-side of Ω in the image), and her being awake is independent from coin toss results. (The image suggests that SB is awake for about 2/3 of the time when she's not under a sleeping spell, but this value is inconsequential.) When SW is undergoing WF's spell, she sleeps most of the time, and even a bit more if the coin landed heads.

The probabilities mentioned in Sect. 2.2 should be interpreted as already conditionalized on S. In addition, Fig. 3 assumes that the duration of one waking period lasts 1/20 of the total duration of the spell, rather than 1/1000 as in Sect. 2.2.

Points of the sample space can be thought of as representing maximally specific descriptions of ways the world could be like:[19] this includes information on what time it is (relative to each agent, if relativistic effects are relevant) and which agents are alive and awake. The partitions coarse-grain this into certain contrasting cases, leaving most details unspecified.

One crucial part of the structure of the Sleeping Beauty problem that this picture leaves out is a partition into time slices, each member of which contains an element of H and an element of ¬H. The members of this partition that intersect with S are equiprobable for a modeller with the same information as SB. This corresponds to random sampling and is crucial for the analogy with Bertrand's boxes problem.

[19] It is here that the terminology of probability theory on the one hand and some branches of physics, logic, and philosophy on the other may come apart. For example, in a probabilistic context, 'the world' may refer to time slices of (part of) the physical universe—whatever is necessary to model the problem at hand.

4.2 Lest we forget

The crux of the memory erasure in the Sleeping Beauty set-up[20] is that it essentially allows the agent to sample random moments from her own lifetime—something which usually can be done only by another agent (e.g., by the prince looking at a randomly picked picture). It is *not* the case that SB randomly samples moments at which she is awake. Although it is admissible to consider a coarse-grained sample space that only represents two possible outcomes (first or second awakening), this does not imply that their probabilities are equal (so the sampling is not random). As was already argued for by the analogy in Sect. 2, the probability associated with the possible waking moments is in fact unequal: it is 2/3 for the first moment versus 1/3 for a second one. Put differently, the relevant evidence SB receives is equivalent to that of a prince receiving a picture taken randomly during the spell, in which it so happens that SB is awake, *not* to that of a prince receiving a picture taken randomly during a moment when SB is awake. Starting from the full sample space in Fig. 3 is helpful to avoid reasoning by this false analogy (*cf*. footnote 7), which leads to the erroneous answer of 1/2 in both Bertrand's boxes problem and the Sleeping Beauty problem.

Observe that solving the Sleeping Beauty problem does not require us to model SB's belief change from before the spell has taken effect to a moment where it has. Elga (2000) and other authors have emphasized that her degree of belief in a fair coin flip resulting heads has to change from one waking moment (right before she's put to sleep) to the next (upon awakening during the spell).[21] Yet, all that is being asked is to model her rational degrees of belief at an awakening during the spell and I have argued that this is structurally analogous to Bertrand's boxes problem. Nevertheless, the puzzle does invite further questions along these lines.

Visualizing the sample space helps us to tease apart which aspects of the Sleeping Beauty problem can or cannot be modelled in an orthodox Bayesian way:

- There is no forgetting involved from just before the sleeping spell till the first waking episode during the spell. It simply ceases to be a moment 'before', but it is now a moment 'during' WF's sleeping spell; this means it is no longer a possible world from ¬S (righthand-side of the sample space in Fig. 3), but rather from S (lefthand-side; *cf*. footnotes 12–14). Although SB does not know that it is the first awakening, an orthodox Bayesian can model it as a normal update.
- There is forgetting involved from the first till the second waking episode (if there is one). Effectively, what this amnesia does is to allow SB to randomly sample waking episodes of her own life time and us to model it in analogy with Bertrand's boxes problem, which can be analysed using Bayes' theorem. It is true that this cannot be modelled by an orthodox Bayesian, who takes evidence to be strictly cumulative.

[20] The comments in this section also apply, mutatis mutandis, to B in the Wicked Joseph variant.

[21] In addition, it is often claimed that this involves forgetting, but this seems wrong: SB does neither forget it is Sunday (rather, it simply ceases to be Sunday, *cf*. Bradley (2011a, b) mutation principle), nor the outcome of the toss (which she was not allowed to see in the first place), nor any details of the procedure (which are crucial for her to know).

– Similarly, there is forgetting from the last waking episode till after the procedure, such that SB doesn't remember how often she has been awake. In addition, it ceases to be 'during', such that the rational degree of belief she should assign to the coin having landed heads is 1/2 (¬S; righthand-side of the sample space in Fig. 3).

So, two things happen simultaneously:

– Time passes and the spell is activated.
– SB is awake.

SB learns these two things (nearly) simultaneously, but the order doesn't matter. Because the spell is activated, she is maximally uncertain about the time (within S). The rational thing to do, looking at the sample space of Fig. 3 is to restrict to the ┐-shaped part on the lefthand-side: the intersection W ⋂ S. This can be made palpable to an orthodox Bayesian by observing that each moment in time is a new one, of which usually little is known in advance. It is because something *is* known in advance (that WF will cast the spell and how it affects SB's awakenings) that probability theory can be applied to it. Probability depends both on our ignorance and on our knowledge, as Laplace already knew.

5 Boxing up the subjectivity in the Sleeping Beauty problem

Winkler (2017, p. 581) gauges the 'thirder' view to be the dominant position regarding the Sleeping Beauty problem, although various minority positions may collectively generate more publications. In mathematics, it can still be interesting to develop a new proof for a well-known theorem. Likewise, in philosophy it is worthwhile to shorten, to clarify, or to optimize otherwise the arguments for a widely shared position. This is exactly what I aimed to do in this paper: to strengthen the defence of the thirder position.

In summary, I have deconstructed the Sleeping Beauty problem into a combinatorial engine and a subjective wrapper. I found the combinatorial engine of the problem to be analogous to Bertrand's boxes paradox, which is solved in terms of standard probability. This core is wrapped in a story asking us to determine a rational degree of belief of a subject, who is in a situation in which her being conscious co-depends on the outcome of a toss with a fair coin. I offered the Snow White problem as a means to see through this outer layer of the Sleeping Beauty problem, independent of its combinatorial core. This shows that subjectivity plays no essential or irreducible role in solving the Sleeping Beauty problem.

The phrasing in terms of subjective probabilities turns out to be merely superficial: the relevant rational degree of belief can be equated with a conditional probability (computable by the usual calculus); it is the latter that does all the work. Since agents cannot observe themselves to be unconscious, subjectivity does introduce a form of selection bias, but it turns out that the probability of interest does not crucially depend on it in this puzzle.

The only way in which subjectivity affects the outcome is that different agents in the story may have different pieces of information: the conditional probabilities

that are relevant to their current situation may have different conditioning events. For instance, the Wicked Fairy knows the outcome of the coin toss, so the probability that she assigns to it having landed heads will be either 0 or 1. But this does not require any new formalism.

It was my aim to solve the Sleeping Beauty problem with the simplest formalism possible, separating the purely combinatorial core from the subjective aspects. An important general lesson is that the sample space has to be made explicit and it is has to be checked that its structure is rich enough for the problem at hand, such that no information has to be added in the form of centered worlds.

Finally, I hope that my analysis of this toy problem can shed some light on more pressing questions pertaining to observer selection effects, for instance in the philosophy of cosmology.

Acknowledgements I am grateful to Igor Douven for proofreading an earlier version of this article and to Jonathan Weisberg and Chris Meacham for invaluable feedback. I thank former student Florijn de Graaf for reviving my interest in this puzzle and audience members present at presentations of parts of this material for encouraging comments. I am grateful to three anonymous referees for their clear comments and suggestions. Part of this research was supported by the Research Foundation Flanders (Fonds Wetenschappelijk Onderzoek, FWO), Grant No. G0B8616N.

References

Bradley, D. J. (2011a). Confirmation in a branching world: The Everett interpretation and Sleeping Beauty. *The British Journal for the Philosophy of Science*, *62*, 323–342.

Bradley, D. J. (2011b). Self-location is no problem for conditionalization. *Synthese*, *182*, 393–411.

Bertrand, J. (1889). *Calcul des probabilités*. Paris: Gauthier-Villars.

Elga, A. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis*, *60*, 143–147.

Halpern, J. Y. (2005). Sleeping Beauty reconsidered: Conditioning and reflection in asynchronous systems. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 1, pp. 111–142). Oxford: Oxford University Press.

Henry, M. (2005). Le problème croix ou pile de d'Alembert, réalité observable et modélisation. In M. Henry (Ed.), *Autour de la modélisation en probabilités* (pp. 219–223). Presses universitaires de Franche-Comté, France (first edition: 2001).

Lewis, D. (1979). Attitudes de dicto and de se. *The Philosophical Review*, *88*, 513–543.

Meacham, C. J. G. (2008). Sleeping beauty and the dynamics of de se beliefs. *Philosophical Studies*, *138*, 245–269.

NWQ. (2011). Archived URL: http://web.archive.org/web/20120111182028/http://www.wetenschap24.nl/programmas/nwq/Answers/Antwoord-14.html. Retrieved Feb 2017.

Rosenthal, J. S. (2009). A mathematical analysis of the Sleeping Beauty problem. *The Mathematical Intelligencer*, *31*, 32–37.

Teller, E., Teller, W., & Talley, W. (2002). *Conversations on the dark secrets of physics*. Perseus Publishing (first edition: 1991).

Titelbaum, M. G. (2013). Ten reasons to care about the Sleeping Beauty problem. *Philosophy Compass*, *8*, 1003–1017.

Titelbaum, M. G. (2014). *Quitting certainties* (2nd ed.). Oxford: Oxford University Press.

Van Fraassen, B. C. (1984). Belief and the will. *Journal of Philosophy*, *81*, 235–256.

Weintraub, R. (2004). Sleeping Beauty: A simple solution. *Analysis*, *64*, 8–10.

Weisberg, J. (2005). Firing squads and fine-tuning: Sober on the design argument. *The British Journal for the Philosophy of Science*, *56*, 809–821.
Winkler, P. (2017). The Sleeping Beauty controversy. *The American Mathematical Monthly*, *124*, 579–587.
Zuboff, A. (1990). One self: The logic of experience. *Inquiry: An Interdisciplinary Journal of Philosophy*, *33*, 39–68.
Zuboff, A. (2009). *Time, self and Sleeping Beauty*. Ph.D. dissertation, Princeton University.