On the Sleeping Beauty problem

January 10, 2013

Suzanne van der Meijden

Institute for History and Foundations of Science

Utrecht University

Supervisor: Professor Dennis Dieks

Foreword

This thesis was written for both the completion of my Bachelor's degree in Physics and Astronomy and my Bachelor's degree in Mathematics at Utrecht University. The research was executed at the Institute for History and Foundations of Science in Utrecht.

I would like to thank the aforementioned institute for providing me with a quiet, yet lively place to work at and, not irrelevant, an unlimited supply of coffee. I want to thank my supervisor, Dennis Dieks, in particular. Without his help, enthusiasm and guidance this thesis would be a far cry from what it is today.

Suzanne van der Meijden,
January, 2012

Abstract

The Sleeping Beauty problem is a well-known problem in self-locating theory. The reason why it is so widely known is that it is a probability theoretic problem where there appear to be two different solutions: 1/2 and 1/3. The main goal of this thesis was to find out how two conflicting solutions can emerge from a mathematical problem and why they both appear to be correct. An attempt at doing this has been undertaken by taking a critical look at the most important publications on the subject. The arguments in these publications have been analyzed extensively by checking whether the claims being made, follow from logical deduction or are simply presumed. This study revealed that several arguments were simply flawed, while others were based on controversial principles. These controversial principles are not mathematical theorems and should therefore be justified by the context of the problem, instead of being presumed at the outset. However, various interpretations of the Sleeping Beauty problem have resulted in divergent conclusions about which principles are applicable in the context of the Sleeping Beauty problem. These contrasting interpretations are a direct consequence of underspecification of the Sleeping Beauty problem: it is not clear what the answer to the problem represents. As soon as one determines what one wants to do with this answer, by adding more context and conditions to the Sleeping Beauty experiment, it is possible to decide which solution is applicable. For example, an analysis on the Doomsday argument, a problem analogous to the Sleeping Beauty problem, showed that the only correct solution in that context is 1/3.

Contents

Introduction

By simply realizing we exist, we can determine the truth of certain statements. For example, if someone claims that you are dead, you can effortlessly find out that this statement is incorrect. Problems that involve one's own existence are called self-locating problems. They are closely connected which suggests that a solution to one of them has considerable influence on the solution to others.

This thesis discusses a very general self-locating problem: the Sleeping Beauty problem. Its universality makes it very useful to investigate, for the solution can enlighten us about similar dilemma's. Many authors have published articles on the subject, the most important among them being from Adam Elga (2000) and David Lewis (2001). What makes the Sleeping Beauty problem so interesting is that different authors come up with different solutions, resulting in discordance.

We want to find out how it is possible that there is no consensus regarding the Sleeping Beauty problem in the scientific community. We will try to answer this question by analyzing the most important arguments proposed.

In the first chapter of this thesis we will introduce the Sleeping Beauty problem and focus on a mathematical description. The core of this thesis consists of an examination of the most important solutions provided so far. We will use our findings to determine the root of the problem. Our final chapter consists of a famous analogous problem: the doomsday argument.

1 The problem

The origin of the Sleeping Beauty problem lies in another self-locating problem: the paradox of the absent-minded driver (Piccione, Rubinstein, 1997). The creators of this paradox constructed many variations of this problem, one of which they could not solve. Arnold Zuboff had independently stumbled on the same matter, but left his findings unpublished. Robert Stalnaker learned of this work and named the problem as we know it today: the Sleeping Beauty problem.

This problem is about an experiment with a subject called Beauty, she is told the following story:

> "Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you belief that the outcome of the coin toss is heads?" (Elga, 2000)

At first sight the problem could appear trivial: before Beauty went to sleep she knew the probability of the coin landing heads was 1/2, when she wakes up she learns nothing new, as she already knew she would be awoken, so she should assign the same credence to heads as before going to sleep: 1/2. Unfortunately, solving the problem is not that simple as one could also argue that there are three possible predicaments in which Beauty could find herself upon awakening: Monday when the coin landed heads, Monday when the coin landed tails and Tuesday when the coin landed tails. In just one of these three cases the coin came up heads, therefore the credence Beauty should assign to heads is 1/3. The group of theorists who think Beauty's credence in heads should be 1/3 are labeled *Thirders* and the other group are called *Halfers*.

1.1 Mathematical description

Probability theory is a mathematical area which is strongly connected to our intuition. When trying to unravel elementary problems this property comes in hand, but when attempting to solve a counter-intuitive paradox it rather inconvenient. In to order avoid falling prey to wrong intuitions, we will only allow ourselves to treat the Sleeping Beauty problem in terms of well-defined events and formulas. Forcing the various halfer and thirder arguments into such a formulation makes them easier to compare and more importantly it shows us whether claims are based on logical deduction only or additional assumptions have sneaked into the argument.

The first step in solving a probability theoretic problem is to define the probability space $\langle \Omega, E, P \rangle$, where the sample space Ω is a complete set of disjoint events, the event space E is the set of all possible generalized events and P represents the probability measure (Uffink, 1990). The Sleeping Beauty experiment commences on Sunday after the coin toss and it ends on Tuesday. For the sake of convenience we will refer to Sunday when Beauty is asleep as Sunday night, Monday when Beauty has been put back to sleep as Monday night and Tuesday when Beauty has been put back to sleep as Tuesday night. During the investigation there are several distinct events, which form the sample space.

Sample space

H_s The coin landed heads, it is Sunday and Beauty is sleeping

T_s The coin landed tails, it is Sunday and Beauty is sleeping

H_m The coin landed heads, it is Monday and Beauty is awake

T_m The coin landed tails, it is Monday and Beauty is awake

T_t The coin landed tails, it is Tuesday and Beauty is awake

S It is Monday or Tuesday and Beauty is sleeping

Thereby, we have $\Omega = \{H_s, T_s, H_m, T_m, T_t, S\}$.

Event space

We see that the sample space is a finite set, which means that we can take the event space as the powerset of these distinct events, thus $E = \mathcal{P}(\Omega)$.

Probability measure

We use Beauty's credence function P as the probability measure. The probability measure is therefore as follows: $P: \mathcal{P}(\Omega) \to [0, 1]$.

1.2 Some terminology

There are some key concepts that will be used throughout this thesis, we will provide a short explanation here. There are possibilities about which world is actual and there are possibilities about one's place in this actual world. The latter possibilities could, for example, regard one's location in time or space and can be represented as classes of so-called *centred* worlds. Possibilities about which world is actual can be represented as classes of *uncentred* worlds (Lewis, 2001).

Let us relate these concepts to our problem: when Beauty awakes she is either in H_m, T_m or T_t, these are examples of centred worlds because these events tell us something about Beauty's place in time. The uncentred worlds are about which world is actual, so there is a world in which the coin landed heads and one where it came up tails.

2 The main arguments

We will open this chapter by summarizing the most important thirder argument, the one provided by Adam Elga. Following we will discuss this reasoning on the basis of remarks made by David Lewis, the creator of the main halfer argument. These remarks will swiftly lead us to his solution to the Sleeping Beauty problem.

2.1 The main thirder argument[1]

Adam Elga opened his famous thirder argument by noticing that when Beauty wakes up there are three possible predicaments she could find herself in: H_m, T_m or T_t. Beauty is in predicament H_m if and only if the outcome of the coin toss is heads. Therefore we can calculate the probability Beauty assigns to being in the heads awakening, to solve the Sleeping Beauty problem. In formula this means that we are looking for:

$$P(H_m|H_m \vee T_m \vee T_t)$$

The difference between T_m and T_t is not a difference in which world is actual, but rather a contrast in Beauty's temporal location. This is because these events both occur in the tails world, but on different moments: T_m takes place on Monday and T_t on Tuesday. As the only asymmetry concerns their place in time, Beauty should assign equal credence to being in each predicament when she learns the outcome of the coin toss was tails, specifically: $P(T_m|T_m \vee T_t) = P(T_t|T_m \vee T_t) = 1/2$. This implies that Beauty assigns equal credence to being in either T_m or T_t upon awakening, in formula:

$$P(T_m|H_m \vee T_m \vee T_t) = P(T_t|H_m \vee T_m \vee T_t)$$

The researchers could execute the experiment a little differently: they could decide to toss the coin on Monday night instead of Sunday night. Regardless of which method was used, the experiment will run the same from Beauty's point of view, therefore her credence function is not modified when the method of coin tossing changes.

Let us assume that the second approach is used. Now, if Beauty is told it is Monday she learns she is either in H_m or T_m. Her credence that she is in H_m is equal to the credence that a fair coin, soon to be tossed, will land heads. This means that she will assign credence 1/2 to being in H_m and thereby also assign credence 1/2 to being in T_m, in formula: $P(H_m|H_m \vee T_m) = P(T_m|H_m \vee T_m) = 1/2$. This implies that Beauty assigns equal credence to T_m and H_m upon awakening, resulting in:

$$P(H_m|H_m \vee T_m \vee T_t) = P(T_m|H_m \vee T_m \vee T_t)$$

[1] This section contains a summary of Elga's argument, a more elaborate explanation can be found in appendix A.

We first found that Beauty assigns equal credence to being in T_m and T_t upon awakening and thereafter showed that she has equal credence in being in T_m and H_m upon awakening. A combination of these two statements immediately implies that Beauty has equal credence in being in either T_m, T_t or H_m upon awakening. Since these credences are equal and sum up to one, we can conclude that Beauty assigns credence 1/3 to being in the heads awakening upon awakening. (Elga, 2000)

2.2 Elga's assumption

Elga's line of reasoning initially appears as pure logical deduction, there is however one claim in Elga's argument which is assumed rather than being deducted. The assumption we are talking about is that Beauty's credence in heads when she is told it is Monday is equal to 1/2, in formula: $P(H_m|H_m \vee T_m) = 1/2$. Elga supports his claim by stating that this credence is equivalent to Beauty's credence in a future coin toss landing heads.

2.2.1 The principal principle

Elga's assumption could be supported by the so-called principal principle, this is not a mathematical axiom but rather a constraint on rational reasoning and therefore depends on the fulfillment of certain conditions to be convincing. This principle was first formulated by Mellor (1971) and put in its final form by Lewis (1980, 1994) and is as follows:

> Credences about future chance events should be equal to the known chances unless there is other relevant information available that should be taken into account.

It is this principle that makes it reasonable to believe that a coin toss in the future will land heads and tails with equal probability. If we were, however, to find out that this coin is not fair, we should adjust our credence to this information and disregard the known chances. Additionally, this principle appears to imply that once Beauty learns it is Monday she assigns degrees of belief to being in either H_m or T_m.

In reaction to Elga's thirder argument, David Lewis pointed out that the principle came with a proviso: one's credences are constrained by one's beliefs about objective chances only if one does not have inadmissible information. According to Lewis, this condition is not satisfied when Beauty is told it is Monday. This is because when she receives this information, she obtains information about her location in time. This news is relevant for Beauty's degree of belief in heads as it now increases[2] from 1/3 to 1/2. From Lewis's point of view, this illustrates that the centred evidence that it is Monday is inadmissible, therefore the proviso applies, invalidating Elga's assumption (Lewis, 2001).

We find Lewis's reasoning not obvious, and we are not alone. According to Nick Bostrom, labeling information as inadmissible is a delicate matter. What counts as inadmissible depends on fulfillment of certain conditions and this should follow from the solution to the problem, rather than being assumed at the outset (Bostrom, 2007). We are thereby unable to decide whether we should accept or reject the usage of the principal principle to justify Elga's assumption.

2.3 The main halfer argument

The main halfer argument has been devised by Lewis and it is closely connected to Lewis's criticism on Elga. Lewis's most important idea is that only new relevant evidence, centred or uncentred, should produce a change in credence and the evidence of waking, $H_m \vee T_m \vee T_t$, is not relevant to the coin landing heads or tails. This means that Beauty's credence in the coin landing heads when she is awoken should be equal to her credence in heads when she went to sleep: 1/2 (Lewis, 2001). This is an application of the principal principle, but now Beauty has to set her credences equal to the known credences in the past, instead of in the future.

2.3.1 Discussion

We take a look at a striking consequence of Lewis's conclusion which he, in fact, pointed out himself. According to Lewis, Beauty's credence in heads when she knows it is Monday is equal to 2/3, and not equal to 1/2 like Elga assumes[3]. Lewis explains this remarkable result by stating that learning that it is Monday is inadmissible, but the information that Beauty is awake, is not.

We think that one could similarly argue that when Beauty awakes she receives the centred evidence that she is awake, which does give her information about her temporal location. She learns that she is currently in one of three possible awakenings, the information that it is Monday narrows these three possibilities down to two. Lewis is in our opinion unable to justify his distinction between relevant and non-relevant evidence and therefore we are unable to decide whether Lewis's reasoning is correct.

[2] According to Elga $P(H_m|H_m \vee T_m) = 1/2$ where $P(H_m|H_m \vee T_m \vee T_t) = 1/3$, we see that learning that it is Monday increases Beauty's degree of belief in heads by one sixth.

[3] A proof of the equivalence of this claim and the halfer position can be found in appendix B.

3 Two variations

In this chapter we will discuss two arguments based on variations on the Sleeping Beauty problem: the first favoring thirders and the second serving halfers.

3.1 A variation in favor of thirders

Cian Dorr has devised a thirder argument using a variation on the original experiment. The modified problem is quoted below:

> "Again, Sleeping Beauty knows for certain on Sunday that she is to be the subject of an experiment. This time, the experimenters will definitely wake her both on Monday and Tuesday, administering an amnesia-inducing drug between the two awakenings. However, they have two amnesia-inducing drugs, and they will decide which one to administer by tossing a fair coin on Monday night. If the outcome of the toss is tails, they will administer the amnesia-inducting drug that was used in the original version of the experiment. If the outcome is heads, they will administer a much weaker amnesia-inducing drug, which merely delays the onset of memories from the previous day, rather than destroying them entirely. If Beauty receives this weaker drug, the first minute of her awakening on Tuesday will be just as it would have been if she had received the stronger drug, but after that the memories of Monday's awakening will come flooding back. She will then realize that it is Tuesday , and that the outcome of the toss must have been heads." (Dorr, 2002)

Dorr now reasons as follows: if Beauty awakes during this new experiment there are four predicaments she could find herself in: H_m,H_t[4], T_m and T_t. Because Beauty is unable to differentiate among these four cases, she assigns equal credence to each. After a while she reaches a new state X in which she is awake long enough for the weaker potion to have lost its effect. If she reaches state X and does not experience memories flooding back, she knows she is not in H_t. The ratio among her other credences should stay equal, and since they sum up to one we obtain $P(H_m|X) = 1/3$. As there are no relevant differences between this case and the original, the thirder position is correct. (Dorr, 2002)

3.1.1 Discussion

Not long after Dorr published his variation, Darren Bradley pointed out that the case Dorr describes is nonequivalent to the original: there is a crucial difference between state X in the variant case and state X in the initial problem. In the

[4] This means that the coin landed heads, it is Tuesday and Beauty is awake.

modified case Beauty can know for sure that she is in H_t due to a flood of memories that reaches her. This implies that the absence of such information makes the coin landing heads less likely. Therefore the ratio among the scenario's H_m, T_m and T_t should not stay equal, contrary Dorr's idea. (Bradley, 2003)

3.2 A variation in favor of halfers

Roger White does not consider himself a halfer, yet he constructs an argument for this party. He does so by posing a challenge for thirders which strongly suggests that the halfer position is correct. His challenge involves the generalized Sleeping Beauty problem quoted below:

> "A random waking device has an adjustable chance $c \in (0, 1]$ of waking Sleeping Beauty when activated on an occasion. In those circumstances in the original story where Beauty was awakened, we now suppose only that this waking device is activated." (White, 2006)

White additionally defines the following events:

H	The coin landed heads
T	The coin landed tails
W	Beauty is awake at least once during the experiment

It is clear that when $c = 1$ the problem is equivalent to the original. However, the $c < 1$ case is significantly different, according to White. This is because in the latter case Beauty is no longer certain about being awoken at all during the experiment. This uncertainty implies that waking up provides Beauty with information regarding the outcome of the toss coin, for she has a greater chance of being awoken if the coin lands tails than if it would have landed heads.

We can see this by calculating the probability of waking given the outcome of the coin toss. When the coin lands heads there is one potential awakening, thus the machine will be activated once, which immediately implies that the probability of being awoken is equal to c. We can thus state:

$$P(W|H) = c$$

When the coin lands tails, the chance of being awoken during the experiment is equal to one minus the probability of not being awoken at all. This means that:

$$P(W|T) = 1 - P(\sim W|T) = 1 - (1 - c)^2 = 2c - c^2$$

We are now ready to apply Bayes's rule:

$$\begin{aligned} P(H|W) &= \frac{P(W|H)P(H)}{P(W|H)P(H) + P(W|T)P(T)} \\ &= \frac{c \cdot \frac{1}{2}}{c \cdot \frac{1}{2} + (2c - c^2) \cdot \frac{1}{2}} \\ &= \frac{1}{3 - c} \end{aligned}$$

We retrieve the original situation by taking $c = 1$, which yields: $P(H|W) = 1/2$. We thereby conclude that 1/2 is the correct answer to the Sleeping Beauty problem. (White, 2006)

3.2.1 Discussion

We have not been able to detect flaws nor incorrect assumptions in White's argument, right up until the last line. We agree that $P(H|W) = 1/2$, but think that this is not the solution to the Sleeping Beauty problem. Recall that we are looking for Beauty's credence in heads upon awakening. White is calculating Beauty's credence in heads when she knows she has been awoken at least once. We think that the information 'Beauty is awake' is nonequivalent to 'Beauty is awoken at least once', which implies that White is calculating a different probability than the one we are looking for.

White's mistake is that he confuses Beauty's perspective with ours. If we learn that Beauty has been awoken at least once, we learn that the coin landing tails is more likely than it landing heads. However, Beauty cannot be told she is never awoken, for she needs to be awake for that information to reach her. The Sleeping Beauty problem is about self-locating belief, and not about Bayesian learning for an outsider, it are these things that White got confused with.

Another important feature of White's variation is that Elga's argument can be constructed again, without being influenced by $c < 1$. Elga's solution is based on two things, the first one is that Beauty assigns equal credence to being in T_m or T_t upon awakening, which is supported by the idea that these two events are part of the same world. The other ingredient is that Beauty assigns equal credence to being in H_m and T_m when she is awake, which follows from Elga's main assumption: $P(H_m|H_m \vee T_m) = 1/2$. This assumption was motivated by the idea that this credence is equal to the probability that a coin, soon to be tossed, will land heads. Both of these ingredients are uninfluenced by $c < 1$, which means that Elga will still find a probability of 1/3. White was well aware of this and took it as a sign that something was wrong with Elga's argument, but we perceive it as a sign that something is wrong with White's.

4 General arguments

In this chapter we will discuss types of arguments that are used by both halfers and thirders. We will start with frequency arguments and close with so-called Dutch strategy arguments.

4.1 Frequency arguments

Frequentism is one of many interpretations of probability theory. According to frequentists, probability is only defined when an experiment can be repeated arbitrarily often. The odds getting a certain outcome are associated with the relative frequency of the occurrence of that outcome in the experiment in a very long series of independent trials (Uffink, 1990). We will not discuss the validity of this interpretation here - for this would the subject of a whole new thesis. Instead, we will devote this section to a thirder and a halfer frequency argument.

Halfer argument

If the Sleeping Beauty experiment is repeated n times we expect $n/2$ head landings and $n/2$ tail landings. Thus the relative frequency of the coin landing heads is equal to:

$$\frac{1/2 \cdot n}{n} = 1/2$$

This means that Beauty should assign credence 1/2 to the coin landing heads.

Thirder argument

Let us take a second look at repeating the experiment n times. As said before, we expect that the coin will land heads $n/2$ times and tails $n/2$ times. This means that Beauty will be asked about her credence on $3n/2$ occasions, where in $n/2$ of these questionings the coin landing heads preceded her awakening. Thus the relative frequency of the coin landing heads when she is asked about her credence is equal to:

$$\frac{1/2 \cdot n}{3/2 \cdot n} = 1/3$$

This means that Beauty should assign credence 1/3 to the coin landing heads.

4.1.1 Discussion

Clearly, thirders and halfers are determining a different frequency. The conflicting interpretations on what frequency we are looking for illustrate that solving the Sleeping Beauty problem is not only a matter of mathematics and assumptions, interpretation plays a key role as well.

There are good arguments for both interpretations: if Beauty would have to guess how the coin landed upon each awakening, she would be correct one third of the time if she guessed heads. However, if we look at the frequency where Beauty guessed correctly in all awakenings during one trial, she would be correct half the time if she went with heads (Armstrong, 2011).

4.2 Dutch book arguments

Dutch book arguments play an important role in the Sleeping Beauty discussion as they are very intuitive. These arguments are based on the idea that a rational person will never accept a bet that will lead to a sure loss. This claim is specified in the so-called converse Dutch book theorem, which is as follows:

> For a collection of betting quotients that obeys the probability axioms, there is no set of bets with those quotients that guarantees a sure loss to one side.

To understand this theorem we need to introduce the concept of a Dutch book: a set of conditions under which a collection of bets guarantees a loss to one side. Such wagers revolve around the truth of a proposition H: one has to pay qS and if H is true, that person receives S. An important requirement for these bets is that they have to be fair, which means they have an expected value of zero using one's betting quotient q. Dutch book arguments rest on the idea that this betting quotient represents credence one has in the truth of proposition H.

A Dutch strategy or diachronic Dutch book is a Dutch book where bets are placed at different moments in time. There has to be a certain algorithm that is available at the outset guaranteeing a profit to one side (Hájek, 2008).

Both halfers and thirders Dutch strategy arguments have been developed. First, we will discuss a halfer argument which was constructed and then criticized by Christopher Hitchcock and thereafter we consider a thirder argument designed by Kai Draper and Joel Pust.

Halfer argument

This halfer argument starts off by showing that the thirder solution cannot be correct, for this will lead to Beauty being Dutch booked. The second part shows that this book can be constructed for any value other than 1/2, implying that Beauty can only assign credence 1/2 to the coin landing heads.

When Beauty follows the thirder line of reasoning, she will consider the following set of bets fair: on Sunday she has to pay \$15 and gets \$30 if the coin landed heads and on Monday she pays \$20 and receives \$30 if the coin landed tails. She will regard these bets as fair because her expectation value of the Sunday and

Monday bets are respectively equal to:

$$1/2 \cdot 30 + 1/2 \cdot 0 - 15 = 0$$
$$2/3 \cdot 30 + 1/3 \cdot 0 - 20 = 0$$

This means that Beauty is willing to pay $\$20 + \$15 = \$35$ but can only win \$30, therefore she will definitely lose \$5. As Beauty is Dutch booked we can state, invoking the converse Dutch book theorem, that thirders cannot be correct.

We can generalize this argument to any other value than 1/2 because the Dutch strategy merely depends on a change in credence. This can be understood by realizing that the shift in credence makes Beauty willing to accept a bet where she has to pay more on Sunday than she can earn on Monday (or vice versa).

Discussion An essential constraint on Dutch book arguments is that there needs to be an algorithm for placing the bets that is available at the outset. This means that the bookie is not allowed to exploit information that is not available to the agent. In the above argument the second bet will take place while the bookie knows it is Monday and Beauty does not. If Beauty would have had the same knowledge as the bookie, she would have updated her belief in heads to 1/2 (this is Elga's main assumption), and thus would not have accepted the second bet as she no longer regards it as fair. This is because the expectation value of this second bet is now negative:

$$1/2 \cdot 30 + 1/2 \cdot 0 - 20 = -5$$

This means that the criteria for a Dutch book are not met and therefore the above halfer argument is invalid (Hitchcock, 2004).

Thirder argument

We will now discuss a Dutch book devised by thirders, this time both Beauty and the bookie will have the same knowledge during the process. The set-up of this argument is generally the same as the above halfer argument.

If Beauty follows the halfer line of reasoning, she will consider the following set of bets fair: on Sunday the bookie sells her a bet which costs \$15 and pays \$30 if the coin landed tails and on Monday, after she and the bookie are told it is Monday, the bookie sells her a second bet which costs \$20 and pays \$30 if the coin landed heads. Beauty considers this set of bets fair because their expectation values are respectively equal to:

$$1/2 \cdot 30 + 1/2 \cdot 0 - 15 = 0$$
$$2/3 \cdot 30 + 1/3 \cdot 0 - 20 = 0$$

Again, we see that Beauty pays \$35 but can only win \$30, resulting in a \$5 loss.

This Dutch strategy can be generalized so that it applies to any value other than 1/3, as it merely depends on the bookie knowing on Sunday that Beauty's credence that the coin will land heads is different from her credence that the coin landed heads if she learns it is Monday. Invoking the converse Dutch book theorem, this means that the thirder solution is the correct one (Draper, Pust, 2007).

4.2.1 Discussion

The above thirder argument satisfies all conditions demanded by the converse Dutch book theorem. Does this mean that we can get ready to celebrate and congratulate the thirder movement? Unfortunately not. Dutch book arguments are very complicated and besides that highly controversial. To illustrate this we will briefly discuss the history of Dutch books regarding the Sleeping Beauty problem.

The first Dutch books regarding the Sleeping Beauty problem were devised by Christopher Hitchcock. He created and criticized the halfer Dutch book described above and put forward a seemingly correct thirder Dutch book (Hitchcock, 2004). Two years after his publication, Darren Bradley and Hannes Leitgeb produced an article criticizing Hitchcock's thirder Dutch book (Bradley, Leitgeb, 2006). Not long after this, Kai Draper and Joel Pust published an article attacking this criticism and additionally presenting new criticism on Hitchcock's article. They also devised the thirder Dutch book described above (Draper, Pust, 2007).

Besides the rough history of Dutch books regarding the Sleeping Beauty problem, there is controversy about the strength of Dutch book arguments in general. This debate is a result of the assumptions being made in such arguments, like valuing money linearly, having betting credences that match betting quotients and the so-called package principle.

Let us pay closer attention to this last principle. Dutch strategy arguments assume that the value of a collection of bets is equal to the sum of the values of the individual bets, this is the package principle (Hájek, 2008). One can find numerous objections against this principle, and the Sleeping Beauty problem can illustrate one of these objections. Consider the thirder Dutch strategy described in the previous section: on the days the bets are placed, Beauty considers the bets fair because their expectation value is, according to Beauty, equal to zero. However, if she considers the betting process as a whole she is able to determine that she will suffer a net loss of \$5. Therefore, Beauty might find the individual bets reasonable to accept, but she will not agree with the collection of bets, implying that Beauty cannot be booked.

5 Root of the problem

So far, we have not come across an argument that has been able to persuade us to join the halfer nor the thirder party. A frequency solution appeared easy, but it turned out not to be as it provided us with two different results. We have seen a flawed halfer and a correct thirder Dutch Strategy argument, unfortunately we cannot use this as a definite proof for there are numerous assumptions involved. In addition, we think that it is possible to construct a correct halfer Dutch Strategy.

On top of the aforementioned arguments we concentrated on two variations of the problem; Dorr's variation formed a thirder argument and White's was meant to pose a challenge for thirders, implicitly serving as a halfer argument. Dorr's reasoning turned out to be invalid, since his version of the problem was nonequivalent to the original story. White's argument turned out to be ineffective as he was not calculating the credence we were looking for. Elga's and Lewis's solutions have not been able to help us any further either, for they both rely on the controversial applicability of the principal principle.

In this chapter we will discuss several opinions on the cause of the unsettled argument between halfers and thirders. We will start with Nick Bostrom who thinks that the differing views on the problem are a consequence of conflicting assumptions in self-locating theory. Next, we will take a look at Berry Groisman's ideas on the problem. According to him, there is no paradox, just confusion about what the event *the coin landed heads* means. Finally, we will discuss a surprising angle provided by Stuart Armstrong, who claims to have solved the problem by finding the correct decision for Beauty to make.

5.1 Self-locating assumptions

Nick Bostrom (2002) relates the halfer and thirder solution to rivaling self-locating assumptions: the Self-Indicating Assumption (SIA) and the Self-Sampling Assumption (SSA). We will define both below and add an application to illustrate the meaning.

Self-Indicating Assumption Given the fact that one exists, one should (other things equal) favor hypotheses according to which many observers exist over hypotheses in which fewer observers exist.

The meaning of this assumption can be best understood by considering the tale of the presumptuous philosopher. Imagine that scientists have narrowed down the search for a theory of everything to no more than two theories. One theory describes a universe containing a *trillion* observers and the alternative involves a universe holding a *trillion trillion* observers. The presumptuous philosopher learns of this dilemma and claims to have found which theory is correct. He argues, using SIA: 'as the second theory describes a universe which contains a trillion times more observers than the universe described in the first theory, this second theory is a trillion times more likely. We can thus safely conclude that the second theory is correct.'

Self-Sampling Assumption All other things equal, an observer should reason as if they are randomly selected from the set of all actually existent observers (past, present and future) in their reference class.

A fun way to apply SSA is to take a look at traffic analysis. When we are on the road or in a supermarket it very often appears as if we are in the slowest lane. We often attribute this inconvenience to psychological effects or even Murphy's Law but there is a much easier explanation: more often than not, the other lane really is faster. This can be understood by realizing that the turtoiselike pace of a queue is often caused by an excess of people using that particular lane. This means that there are more people in a sluggish lane, than in a fast one. By simply appealing to SSA this means that there is a greater prior probability of being in a slow lane than in a fast one.

Bostrom has finally developed a combination of SIA and SSA which is as follows:

SSA + SIA All other things equal, an observer should reason as if they are randomly selected from the set of all possible observers.

5.1.1 Relation to the Sleeping Beauty problem

Now that we know the various self-locating assumptions, we are ready to relate them to our problem. According to Bostrom the thirder solution follows from a combination of SSA and SIA and the halfer solution follows from applying SSA. We can easily see why:

Thirder solution

We know that in the Sleeping Beauty problem, a heads landing will create one Beauty awakening and a tails landing will create two. This means that in total there are three possible observers, each existing with probability 1/2, so SIA+SSA assigns 1/3 probability to each.

Halfer solution

In the Sleeping Beauty problem, there are two worlds that can be created: a heads world with one observer or a tails world with two observers. As this depends on the toss of a fair coin these worlds are equally probable, hence the SSA probability of being the first (and only) observer in the heads world is 1/2.

5.1.2 Discussion

Bostrom is neither a halfer nor a thirder, specifically, he thinks both answers are incorrect. He argues that the halfer and the thirder line of reasoning will generate unacceptable consequences and are therefore false (Bostrom, 2007). Bostrom provides an alternative solution that combines the desirable intuitive properties of the halfer and thirder solution, based on a self-developed self-locating observation theory. We cannot discuss this theory and the corresponding solution to the Sleeping Beauty problem here, but if one is interested it can be found in his book: *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (2002).

We question Bostroms motives to reject the halfer and thirder solution for he uses *reductio ad absurdum* to do so. Showing that a certain proposition has some remarkable, and perhaps even counter-intuitive consequences is not a formal way of showing that something is incorrect. I do, however, agree on his stand that SIA+SSA implies the thirder solution and SSA implies the halfer solution.

5.2 Different interpretations

Groisman intends to work out the Sleeping Beauty problem by arguing that thirders and halfers are interpreting the problem differently. He identifies the improper use of the notion of an event as the origin of these divergent views. According to Groisman, we need to add an experimental set-up to describe the event the coin landed heads. In the question posed to Beauty we implicitly assume the set-up. We are not simply asking Beauty about her credence in the coin landing heads, but rather:

> What is your credence that the coin landed heads under the set-up of awakening?

Which is equivalent to:

> What is your credence that this awakening is a heads awakening under the set-up of awakening?

Groisman argues that one could also interpret the initial question as:

> What is your credence that the coin landed heads under the set-up of coin tossing?

According to Groisman, the answer to the first and the second question is obviously 1/3 whereas the answer to the third question should be 1/2. He concludes that the root of the problem is that one tends to confuse the different questions, thereby arriving at an apparent paradox (Groisman, 2008).

5.2.1 Discussion

Groisman argues that halfers have mistaken what the Sleeping Beauty problem is about. He assumes that halfers are trying to answer a question that is not actually posed to Beauty. This criticism is erroneous as halfers are just as much trying to find Beauty's credence in heads under the set-up of awakening. Lewis has even explicitly mentioned this in his 2001 article.

Despite that, his ideas on the root of the problem being more of a interpretation issue rather than a mathematical issue are in our view correct. The way Groisman puts forward his ideas is a bit clumsy, but there definitely is an issue with the notion of an event, which we already encountered in the differing frequency solutions. We think the problem is not the event itself, but different ideas on what reference class the event belongs to. These ideas are in line with Bostroms ideas on the subject, suggesting that the reference classes in SIA and SSA form the problem.

5.3 A decision theoretic perspective

We will now discuss Stuart Armstrong's (2011) take on the root of the problem. He has a very different approach to the Sleeping Beauty problem and comes up with a surprising conclusion. According to Armstrong the problem is underspecified from a decision theoretic perspective which results in the differing intuitions on the matter. In order to overcome this issue he sets out to solve the problem, not by calculating a probability but by finding the correct decision to make.

The Sleeping Beauty problem as it stands, does not involve Beauty making decisions. Therefore, we assume that whenever Beauty is awoken she is offered a coupon that pays out \$1 if the coin landed tails and 0\$ if it landed heads. She must decide at what cost she is willing to buy the coupon. The amount she is willing to pay depends on what kind of person Beauty is: she could be either selfless, altruistic or selfish on one hand and on the other hand either a total or an average utilitarian. Now, assuming that Beauty's utility function is linear in cash, we are ready to analyze four cases.

Selfless Sleeping Beauty Selfless Beauty follows non-personal preferences that do not include any other agent's personal preferences. In the tails world, Beauty's future self will be offered the same deal twice, meaning that every profit that she makes will be doubled. This means that if Beauty's future 'copies' buy the coupon for $\$x$, she expects to earn:

$$\$0.5\,(2 \times (1-x) + 1 \times (0-x)) = \$(1 - 3/2 \cdot x)$$

This profit is positive for $x < \$2/3$, so for that price Beauty would want her future copies to buy a coupon.

Altruistic total utilitarian Sleeping Beauty An altruistic Beauty does take other agent's preferences into account when making decisions, contrary to selfless Beauty. However, if this altruistic Beauty is a total utilitarian she will make exactly the same decision as selfless Beauty. This is because profits in the tails world are doubled, as any gain or loss will happen twice. An altruist simply adds up the effects of these gains or losses and will thereby arrive at the same solution as selfless Beauty. This means that an altruistic total utilitarian Beauty will advice her future copies to buy a coupon if the price is less than \$2/3.

Altruistic average utilitarian Sleeping Beauty The case for an altruistic average utilitarian Beauty is significantly different. If Beauty's future copies decide to pay $\$x$ for the coupon, they will make $-\$x$ in the heads world and $\$(1-x)$ in the tails world, per copy. This means that each copy expects to earn:

$$\$0.5(-x+(1-x)) = \$(0.5-x)$$

This is positive for $x < \$0.5$, which implies that for that price Beauty would advice a future copy to buy the coupon.

Selfish sleeping Beauty A selfish Beauty does not care about what her past nor future self will gain, she only cares about *hic et nunc*. As every Beauty wants the maximal gain for herself, a copy will reason as follows: buying a coupon for $\$x$ yields $-\$x$ in the heads world and $\$(1-x)$ in the tails world, meaning that I expect to earn:

$$\$0.5(-x+(1-x)) = \$(0.5-x)$$

This means that a copy will buy the coupon for $x < \$0.5$, just like in the altruistic average utilitarian case.

5.3.1 Discussion

What is so striking about the above findings is that a selfless or total utilitarian Beauty will make thirder-like decisions whereas a selfish or altruistic average utilitarian Beauty will make halfer-like decisions. We think that these differing decisions reveal that the problem is underspecified, which probably is the cause for the differing solutions to the Sleeping Beauty problem.

6 Doomsday

We will close this thesis with a problem that is very closely connected to the Sleeping Beauty problem. This analogy was first discovered by Dieks (2007) and we will discuss is paper on the matter below. The doomsday argument presents a line of reasoning suggesting that one should fear doom, for it is near.

6.1 The Doomsday Argument

The doomsday argument differentiates between two scenario's regarding the continuance of human life: the first involves a human race that will perish soon, which is commonly referred to as *doom soon.* The other, *doom late*, concerns human life sustaining for a very long time. Suppose one has determined one's degrees of belief in these two hypotheses, but has forgotten to take into account that one is living now. Without a doubt this is important information, as living now would make one a typical human being in the doom soon scenario and a very untypical one in the doom late scenario. For in the latter situation there would live many more people over time and thus the probability of being alive *now* is very small. We can conclude that updating one's beliefs on the information that one lives now increases one's credence in doom soon and thereby decreases one's credence in doom late.

We will now take a more formal look at the doomsday argument by using mathematical description of the problem. We introduce the following abbreviations: D_s represents doom soon, $p_s = P(D_s)$ is the prior probability assigned to doom soon, N_s is the number of people that will live over time if doom soon is true, E_s represents the evidence that one lives before the date of doom soon. Replacing the subscripts s by l gives the abbreviations for the doom late scenario. Additionally, n is the number of people that are living now and E is the evidence of learning that one lives now. Finally, $q_s = P(D_s|E_s)$ and $q_l = P(D_l|E_s)$.

We are looking for the probability of doom soon being true when taking into account that one lives now: $P(D_s|E)$. An easy way to calculate this probability is to use Bayes's rule, which yields:

$$P(D_s|E) = \frac{P(E|D_s)P(D_s)}{P(E|D_s)P(D_s) + P(E|D_l)P(D_l)}$$

In order to find this probability, we first need to determine $P(E|D_s)$ and $P(E|D_l)$, which are easily computed using indifference. The probability of being one of the n people living now is simply n/N_s in the doom soon scenario and n/N_l in the doom late scenario.

We can now calculate $P(D_s|E)$ by filling in the previous equation:

$$\begin{aligned} P(D_s|E) &= \frac{\frac{n}{N_s}\cdot p_s}{\frac{n}{N_s}\cdot p_s + \frac{n}{N_l}\cdot p_l} \\ &= \frac{p_s}{p_s + \frac{p_l N_s}{N_l}} \end{aligned}$$

There will live considerable more people over time in the doom late scenario than in the doom soon situation, which means that we can safely state: $N_l \gg N_s$. If we take this limit of the above formula we will end up with a shocking result. As $N_l \gg N_s$ the following fraction will become zero $\frac{p_l N_s}{N_l} \to 0$, which means that $P(D_s|E) \to 1$. This means that one should think that doom is imminent, independently of the values of one's priors. Perhaps the most remarkable about this result is that we do not think that doom will strike soon, yet our calculations show otherwise. We can only conclude that something has to be flawed in the doomsday argument.

In his 2007 article Dieks solved this mystery; assuming that one has not already taken into account one's current location in time when calculating one's priors, p_s and p_l, is incorrect. This claim can be strengthened by demonstrating how p_s can be found. We do so by splitting p_s into two conditional probabilities, one of which will be zero:

$$\begin{aligned} p_s &= P(D_s) \\ &= P(D_s|E_s)P(E_s) + P(D_s|\sim E_s)P(\sim E_s) \\ &= P(D_s|E_s)P(E_s) \\ &= q_s P(E_s) \end{aligned}$$

Of course, $\sim E_s$ represents the complement of E_s which therefore represents the information of living after the date of doom soon. Obviously, doom soon cannot be true if one lives after the date of doom soon, which implies $P(D_s|\sim E_s) = 0$.

We are left with one unknown in our formula: $P(E_s)$. We can find this probability again using conditionalizing:

$$\begin{aligned} P(E_s) &= P(E_s|D_s)P(D_s) + P(E_s|D_l)P(D_l) \\ &= 1 \cdot p_s + N_s/N_l \cdot p_l \end{aligned}$$

where we used the fact that $P(E_s|D_s) = 1$ and $P(E_s|D_l) = N_s/N_l$. These are logical assumptions because when one learns that doom soon is true, one has to live before the date of doom soon, implying: $P(E_s|D_s) = 1$. We can justify

$P(E_s|D_l) = N_s/N_l$ by realizing that learning that doom late is true, means that one is part of N_l people. E_s represents living before the date of doom soon, which implies that living before this date holds that one will be one of N_s people, which means that $P(E_s|D_l) = N_s/N_l$.

Summarizing the above yields[5]:

$$p_s = \frac{q_s \cdot N_s}{q_s \cdot N_s + q_l \cdot N_l}$$

We can interpret $q_s = P(D_s|E_s)$ and $q_l = P(D_l|E_s)$ as the probabilities one assigns to the doom soon and doom late scenario in full awareness that one lives before doom soon. If one no longer knows one's place in time, one will assign probabilities p_s and p_l to doom soon and doom late. Not knowing one's place in time means that it is possible that one lives after the date of doom soon, which means that one should assign zero probability to doom soon being true. Conditionalizing on the evidence about which year it is, by means of Bayes's rule with p_s and p_l as prior probability will bring one back to q_s and q_s, which one had assigned in the first place.

The crux is that it is inconsistent to assign the same probabilities to hypothesis about what is going to happen after a certain date both in the situation in which one does not know one's place in history and in one's actual situation, in which one knows that the events in question have not yet occurred. This is what makes the doomsday argument invalid.

6.2 Connection to Sleeping Beauty

The issues faced in the doomsday argument are in some respects interchangeable with the difficulties encountered in the Sleeping Beauty problem. When Beauty wakes up she is asked to assign probabilities to the coin having landed heads or tails. Which is equivalent to assigning credences to being awoken either once or twice. If she learns that her current place in time is Monday, the probability she assigns to the coin having landed tails is equal to the probability of being awoken tomorrow.

The analogy with doomsday is self-explanatory. We can thus apply the formula for p_s in pursuance of finding Beauty's credence in heads upon awakening. Recall that a thirder thinks that learning that it is Monday causes Beauty to have credence 1/2 in heads, so we can say that $q_s = P(D_s|E_s) = 1/2$, whereas halfers think that $q_s = P(D_s|E_s) = 2/3$. Finally, we notice that $N_s = 1$ and $N_l = 2$.

We can now fill out the formula for p_s for thirders and halfers:

$$p_s = \frac{q_s \cdot N_s}{q_s \cdot N_s + q_l \cdot N_l}$$

[5] A proof can be found in Appendix C.

We see that we obtain the halfer and thirder solution:

$$\begin{aligned} thirder\ p_s &= \frac{1/2 \cdot 1}{1/2 \cdot 1 + 1/2 \cdot 2} = 1/3 \\ halfer\ p_s &= \frac{2/3 \cdot 1}{2/3 \cdot 1 + 1/3 \cdot 2} = \frac{2}{2+2} = 1/2 \end{aligned}$$

There is a considerable correspondence between the doomsday argument and the halfer line of reasoning. Stating that Beauty's credence in heads should be equal on both Monday and Sunday is equivalent to saying that one should use one's usual probabilities even if one forgets about one's location in time in the doomsday argument. It is this type of reasoning that lead to the conclusion that doomsday is near, which was incorrect. This means that only the thirder solution is applicable in this context.

Conclusion

We started this thesis to identify why two different solutions to the Sleeping Beauty problem appear correct. To reach this goal, we analyzed several arguments presented by both halfers and thirders. These were originally constructed using differing probability spaces, which made them hard to compare. Therefore, we created a more general probability space in such a way that it would be applicable to all arguments.

Our first analysis was on the two most important arguments in the Sleeping Beauty discussion, the ones created by Elga and Lewis. We were unable to determine which solution is correct, as they both rely on contested applicability of the principal principle. Thereafter, we examined two variations on the original experiment. The one proposed by Dorr turned out to be nonequivalent to the original and White's argument is flawed as it assesses a different probability than the one we are looking for. Finally, we discussed argument types appealed to by both groups: frequency and Dutch book arguments. The frequency arguments revealed that halfers and thirders have a conflicting interpretation of the problem. Additionally, we found a correct Dutch book argument in favor of thirders. However, such arguments are highly controversial and cannot be seen as a definite proof.

After having considered these arguments, we were able determine the cause of the disagreement. We did so on the basis of remarks made by Bostrom, Groisman and Armstrong. According to Bostrom, the halfer and thirder solution follow from applying different self-locating assumptions, both are not a-priori correct and should be justified by context. Groisman stated that halfers interpret the problem incorrectly, resulting in an erroneous solution. Finally, Armstrong revealed an underspecification from a decision theoretic perspective. It is this discovery that convinced us that the problem could be interpret in different ways, like Groisman argued. These contrasting interpretations result in different ideas regarding which self-locating assumption is applicable, which in turn leads to either the halfer or the thirder solution being correct.

We conclude that there is no definite answer to Sleeping Beauty problem because it allows multiple interpretations. In some contexts the thirder solution is applicable and in others the halfer solution is. We illustrated this idea by showing that the halfer line of reasoning coincides with the doomsday argument. The doomsday setting is analogous to the Sleeping Beauty experiment, but due to the extra context, only the thirder solution is correct.

There is another problem considered to be analogous to the Sleeping Beauty problem. According to various authors, being a thirder implies that the quantum many worlds interpretation cannot be correct. We will not discuss this statement, but recommend studying Peter Lewis's *Quantum Sleeping Beauty* (2007) for further research.

On top of that, I challenge the reader to construct a Dutch strategy against thirders, for it would give significant extra strength to the halfer position.

To end, we have concluded that both answers to the Sleeping Beauty problem are correct, depending on the chosen context. However, the Sleeping Beauty discussion is still alive and kicking, arguments for both sides are still being put out regularly. These arguments are often based on a connection between the Sleeping Beauty problem and another famous problem. We think that such arguments[6] add strength to our conclusion rather than weakening it. This is because the connection to another problem adds context, resulting in either the halfer or the thirder solution being true.

[6] See for example *Judy Benjamin is a Sleeping Beauty* (Bovens, 2010) or *Confirmation in a Branching World: The Everett Interpretation and Sleeping Beauty* (Bradley, 2011).

References

[1] Armstrong, S. 2011. Anthropic decision theory for self-locating beliefs. *[Unpublished]*

[2] Bovens, L. 2010. Judy Benjamin is a Sleeping Beauty. *Analysis* 70: 23-26

[3] Bostrom, N. 2007. Sleeping beauty and self-location: A hybrid model. *Synthese* 157: 59-78

[4] Bostrom, N. 2002. *Anthropic Bias: Observation Selection Effects in Science and Philosophy.* Routledge: New York.

[5] Bradley, D. 2003. Sleeping Beauty: a note on Dorr's argument for 1/3. *Analysis* 63: 266-268

[6] Bradley, D. 2011. Confirmation in a Branching World: The Everett Interpretation and Sleeping Beauty. *British Journal for the Philosophy of Science* 62: 323-342

[7] Bradley, D., & Leitgeb, H. 2006. When betting odds and credences come apart: More worries for Dutch book arguments. *Analysis* 66: 119-127

[8] Dieks, D. 2007. Reasoning about the future: Doom and Beauty. *Synthese* 156: 427-439

[9] Dorr, C. 2002. Sleeping Beauty: in defense of Elga. *Analysis* 62: 292-296

[10] Draper, K. and Pust, J. 2007. Diachronic Dutch Books and Sleeping Beauty. *Synthese* 164: 281-287

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

[12] Groisman, B. 2008. The end of Sleeping Beauty's nightmare. *British Journal for the Philosophy of Science* 59: 409-416

[13] Hájek, A. 2008. *Dutch Book arguments.* In Paul Anand, Prasanta Pattanaik & Clemens Puppe (eds.), The Oxford Handbook of Rational and Social Choice. Oxford University Press.

[14] Hitchcock, C. 2004. Beauty and the bets. *Synthese* 139: 405-420

[15] Lewis, D. 1980. A subjectivist guide to objective change. In R.C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability.* University of California Press: Berkely.

[16] Lewis, D. 1994. Humean supervenience debugged. *Mind* 103: 473-490.

[17] Lewis, D. 2001. Sleeping Beauty: reply to Elga. *Analysis* 61: 171-176

[18] Lewis, P.J. 2007. Quantum Sleeping Beauty. *Analysis* 67: 59-65

[19] Mellor, H. 1971. *The Matter of Chance.* Cambridge University Press: Cambridge.

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

[21] Uffink, J. 1990. *Grondslagen van het Waarschijnlijkheidsbegrip.* Utrecht University.

[22] van Fraassen, B. 1984. Belief and the will. *Journal of Philosophy* 81: 235-256.

[23] White, R. 2006. The generalized Sleeping Beauty problem: a challenge for thirders. *Analysis* 66: 114-119

A Elga

In this section we will discuss Elga's solution of the Sleeping Beauty problem. We will do so by thoroughly analyzing every step in his argument: we will start with the interpretation on the problem and we will then figure out the mathematical details of Elga's argument.

Interpretation

Elga's argument starts off with pointing out that we are looking for the probability of being in the heads awakening, specifically: $P(H_m|H_m \vee T_m \vee T_t)$. Elga supports this claim by stating that upon awakening Beauty will find herself in H_m if and only if the coin landed heads. This claim is of course true and we can see this by translating the Sleeping Beauty problem into formula's. In order to do this we need to determine what *the coin landed heads* and *credence upon awakening* mean.

The event that the coin landed heads is a union of all moments where the coin landed heads: $H \vee H_m \vee S_h$, where S_h subset of S where the coin landed heads. The second step is to determine how to translate credence upon awakening into mathematical terms. We translate this by conditioning Beauty's credence function on the information that she is awake: $H_m \vee T_m \vee T_t$. This results in the idea that Beauty's credence that the coin landed heads upon awakening is equal to $P(H \vee H_m \vee S_h|H_m \vee T_m \vee T_t)$, which can be rewritten as:

$$P(H|H_m \vee T_m \vee T_t) + P(H_m|H_m \vee T_m \vee T_t) + P(S_h|H_m \vee T_m \vee T_t)$$

The events H and S are taking place while Beauty is sleeping, so her credence in these propositions will be zero when she is awake. We see that this implies that we only have to calculate the probability of being in the heads awakening upon awakening to find the probability that the coin landed heads. We are thus looking for $P(H_m|H_m \vee T_m \vee T_t)$ in the Sleeping Beauty problem.

Mathematics

After identifying the problem Elga argues that Beauty should have equal credence in T_m and T_t when she finds out the coin landed tails. Specifically: $P(T_m|T_m \vee T_t) = P(T_t|T_m \vee T_t)$. We agree with this claim because the events T_m and T_t are part of the same actual world, which implies that there are no relevant differences from Beauty's perspective regarding the two predicaments and therefore Beauty will assign equal credence to T_m and T_t when she knows the coin landed tails.

From this Elga concludes that Beauty has equal credence in T_m and T_t upon awakening, in formula:

$$P(T_m|T_m \vee T_t) = P(T_t|T_m \vee T_t) \Rightarrow P(T_m|H_m \vee T_m \vee T_t) = P(T_t|H_m \vee T_m \vee T_t)$$

Proof. We can prove this using Bayes's rule twice:

$$\begin{array}{lll} & P(T_m|T_m \vee T_t) & = P(T_t|T_m \vee T_t) \\ \Leftrightarrow & \frac{P(T_m \vee T_t|T_m)P(T_m)}{P(T_m \vee T_t)} & = \dfrac{P(T_m \vee T_t)|T_t)P(T_t)}{P(T_m \vee T_t)} \\ \Leftrightarrow & P(T_t) & = P(T_m) \\ \Leftrightarrow & \frac{P(T_m)}{P(H_m \vee T_m \vee T_t)} & = \dfrac{P(T_t)}{P(H_m \vee T_m \vee T_t)} \\ \Leftrightarrow & \frac{P(T_m)P(H_m \vee T_m \vee T_t|T_m)}{P(H_m \vee T_m \vee T_t)} & = \dfrac{P(T_t)P(H_m \vee T_m \vee T_t|T_t)}{P(H_m \vee T_m \vee T_t)} \\ \Leftrightarrow & P(T_m|H_m \vee T_m \vee T_t) & = P(T_t|H_m \vee T_m \vee T_t) \end{array}$$

□

Elga continues his argument by stating that since the researchers could toss the coin on Monday night instead of Sunday night, Beauty should have equal credence in heads and tails when she is told it is Monday. He supports this claim by arguing that her credence in heads when she is told it is Monday reflects her degree of belief that a coin, soon to be tossed, will land heads. This statement comes off as fairly reasonable, but it does not follow mathematically from the previous statements. For now we will assume that it is true, thus we accept: $P(H_m|H_m \vee T_m) = 1/2$.

From this assumption Elga correctly concludes that Beauty should have equal credence in H_m and T_m upon awakening, in formula:

$$P(H_m|H_m \vee T_m) = 1/2 \Rightarrow P(H_m|H_m \vee T_m \vee T_t) = P(T_m|H_m \vee T_m \vee T_t)$$

Proof. Since $P(H_m|H_m \vee T_m) = 1/2$ it follows immediately that $P(T_m|H_m \vee T_m) = 1 - P(H_m|H_m \vee T_m) = 1/2$. This means that $P(H_m|H_m \vee T_m) = P(T_m|H_m \vee T_m)$. Now, replacing T_t in the previous proof by H_m gives the desired proof. □

Finally, Elga concludes that as Beauty has equal credence in H_m, T_m and T_m upon awakening and these credences sum up to one, that Beauty should assign probability 1/3 to each predicament. This means that $P(H_m|H_m \vee T_m \vee T_t) = 1/3$.

Proof. From the previous statements we can conclude that $P(H_m|H_m \vee T_m \vee T_t) = P(T_m|H_m \vee T_m \vee T_t) = P(T_t|H_m \vee T_m \vee T_t)$, if we add these credences, which we can do since these events are disjoint, we get $P(H_m \vee T_m \vee T_t|H_m \vee T_m \vee T_t) = 1$. Since these chances are equal and sum up to one, each probability must be equal to 1/3. We thereby arrive at the thirder solution. □

B Lewis

Claim. $P(H_m|H_m \vee T_m \vee T_t) = 1/2 \Rightarrow P(H_m|H_m \vee T_m) = 2/3$

Proof. $P(H_m|H_m \vee T_m \vee T_t) = 1/2$ immediately implies that $P(T_m \vee T_t|H_m \vee T_m \vee T_t) = 1 - P(H_m|H_m \vee T_m \vee T_t) = 1/2$. Since Beauty is unable to distinguish between T_m and T_t and she should assign equal credence to each: $P(T_m|H_m \vee T_m \vee T_t) = P(T_t|H_m \vee T_m \vee T_t) = 1/4$. We are now able to compute $P(H_m|H_m \vee T_m)$:

$$\begin{aligned}
& P(H_m|H_m \vee T_m) \\
&= \frac{P(H_m|H_m \vee T_m)P(H_m)}{P(H_m \vee T_m)} \\
&= \frac{P(H_m)}{P(H_m) + P(T_m)} \\
&= \frac{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{P(H_m \vee T_m \vee T_t|H_m)} \\
&\quad \times \frac{1}{\frac{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{P(H_m \vee T_m \vee T_t|H_m)} + \frac{P(T_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{P(H_m \vee T_m \vee T_t|T_m)}} \\
&= \frac{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{1} \\
&\quad \times \frac{1}{\frac{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{1} + \frac{P(T_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{1}} \\
&= P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t) \\
&\quad \times \frac{1}{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t) + P(T_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)} \\
&= \frac{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)}{P(H_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t) + P(T_m|H_m \vee T_m \vee T_t) \cdot P(H_m \vee T_m \vee T_t)} \\
&= \frac{P(H_m|H_m \vee T_m \vee T_t)}{P(H_m|H_m \vee T_m \vee T_t) + P(T_m|H_m \vee T_m \vee T_t)} \\
&= \frac{1/2}{1/2 + 1/4} \\
&= \frac{2}{3}
\end{aligned}$$

□

C Doomsday

Claim. $\begin{cases} p_s = P(D_s|E_s)P(E_s) \\ P(E_s) = 1 \cdot p_s + N_s/N_l \cdot p_l \end{cases} \quad \Rightarrow p_s = \frac{q_s \cdot N_s}{q_s \cdot N_s + q_l \cdot N_l}$

Proof. We first fill in what we know:

$$\begin{aligned} p_s &= P(D_s|E_s)P(E_s) \\ &= q_s \cdot [p_s + N_s/N_l \cdot p_l] \\ &= q_s p_s + q_s N_s/N_l \cdot (1 - p_s) \\ &= q_s p_s + q_s \cdot N_s/N_l - q_s p_s \cdot N_s/N_l \end{aligned}$$

Rearranging the equation such that all terms containing p_s are on the left side yields:

$$\begin{aligned} p_s - q_s p_s + q_s p_s \cdot N_s/N_l &= q_s N_s/N_l \\ p_s(1 - q_s + q_s N_s/N_l) &= q_s N_s/N_l \end{aligned}$$

We can now extract p_s:

$$\begin{aligned} p_s &= \frac{q_s N_s/N_l}{(1 - q_s + q_s N_s/N_l)} \\ &= \frac{N_s q_s}{q_l N_l + q_s N_s} \end{aligned}$$

□