SLEEPING BEAUTY MEETS MONDAY

KARL KARLANDER LEVI SPECTRE

Section I: THE SLEEPING BEAUTY PARADOX

A. Elga (2000) is responsible for introducing a problem - the Sleeping Beauty Problem - that has generated considerable interest and controversy.[1] The puzzling feature of the problem is that it presents a situation in which there seem to be compelling reasons for accepting contradicting answers. Elga formulates the problem as follows:

> 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 believe that the outcome of the coin toss is Heads?
>
> (Elga, 2000, p. 143)

It is understood that you[2] are informed of the experiment setup on Sunday prior to being put to sleep, that on Monday you will be woken up whether or not the coin lands Heads or Tails, and that on Tuesday you will be woken up if and only if the toss resulted in Tails. With this knowledge in place it seems that one has gained no new information upon being first awakened. One knew, after all, already on Sunday that one would be

[1] It first appeared in Piccione, M. and A. Rubenstein (1997).
[2] We shift freely between second person singular and third person singular. The context should (we hope) facilitate clarity. In any event, no special meaning should be attributed to these shifts.

woken up at least once. One's degree of credence in the toss resulting in Heads should, therefore, be the same as on Sunday, namely, one half. Elga, however, argues that the probability assignment should be one third. He gives two reasons. First, if the experiment were to be repeated a large number of times, roughly one third of the awakenings would be associated with Heads.[3]

The second argument Elga presents rests, first, on the observation that the probabilities of the wakening day being Monday conditional on the coin landing Tails should be the same as the probability of its being Tuesday conditional on Tails: Pr(*Monday*|*Tails*)=Pr(*Tuesday*|*Tails*). This is based on an indifference principle; the two awakenings are subjectively indistinguishable and should therefore be accorded the same credence.[4] By the definition for conditional probabilities, this means that $\frac{\Pr(Monday \wedge Tails)}{\Pr(Tails)} = \frac{\Pr(Tuesday \wedge Tails)}{\Pr(Tails)}$ and thus Pr(*Monday*∧*Tails*)= Pr(*Tuesday*∧*Tails*). Next, Elga notes that if one were to learn that the wakening day is a Monday, one should assign equal probabilities to Heads and Tails. The coin could just as well be tossed on Monday evening, since the action which the toss governs does not take place until then. So what is at issue is the future toss of a fair coin. What this means is that Pr(*Heads*|*Monday*)=Pr(*Tails*|*Monday*) and by the conditional probability definition, we have, $\frac{\Pr(Heads \wedge Monday)}{\Pr(Monday)} = \frac{\Pr(Tails \wedge Monday)}{\Pr(Monday)}$ from which it follows that Pr(*Heads*∧*Monday*)=Pr(*Tails*∧*Monday*). Putting the equations together yields

[3] For reasons to think that this argument is not as compelling as it may first seem, see F. Arntzenius' helpful discussion in Arntzenius (2002).

[4] Lewis (2001: 172) accepts Elga's *restricted* indifference principle prescribing the division of cadences between indistinguishable centers of the same possible world. For some consequences of this indifference principle see Elga (2004).

$\Pr(Heads \wedge Monday) = \Pr(Tails \wedge Monday) = \Pr(Tuesday \wedge Tails)$. But from the perspective of the awakening these probabilities are jointly exhaustive and mutually exclusive, and must therefore all equal one third. Since the first one represents the probability of Heads, the required result follows.

The majority opinion in the debate is that one third is the correct answer. There are some dissenters, though, notably David Lewis, who favors a one half answer. Before proceeding, let us observe that there is a very counterintuitive consequence of the one half position. Lewis, noting the difficulty, does not, however, consider it to be unacceptable (2001: 175). Let us look at the credence that Sleeping Beauty should assign to the Heads hypothesis upon being told, sometime during the first day, before the toss, that it is Monday (and after considering his credence about the toss, as Lewis advises). This probability is: $\Pr(Heads|Monday) = \Pr(Monday|Heads) \times \frac{\Pr(Heads)}{\Pr(Monday)}$. Consider the probabilities on the right hand side of the equation. If Sleeping Beauty were to learn - somehow - upon awakening that the result of the toss would be Heads, he would know that it is Monday. So $\Pr(Monday|Heads)=1$. Furthermore we may assume that he does not, in fact, know upon awakening that it is Monday, so that $\Pr(Monday)<1$ and the prior probability of Heads is half. But this means that the proponent of the one half answer is committed to the claim that $\Pr(Heads|Monday)>½$. In other words, Sleeping Beauty assigns a probability greater than one half to the outcome of Heads of a future toss of what he knows to be a fair coin.

This problem generalizes. Any account which entails a probability for the day being Monday (before learning that it is Monday) that is not twice as high as the probability of Heads, will face the same predicament when conditionalizing on new information that it

is Monday.

Elga's one-third view avoids the above difficulty. But it has problems of its own. Lewis and Elga share the assumption that between Sunday and Monday after Sleeping Beauty is awakened, he does not gain any new information or evidence (at least nothing relevant). This leads Lewis to conclude that the probability remains the same and Elga to admit that the shift in credence is strange (2000: 145-6). But in fact things are worse than merely strange on Elga's view. By considering the following principle we can see this more clearly.

> No New Evidence Principle: One has warrant for changing one's credence in the truth of a proposition only if one has received new relevant evidence or information (including evidence about the coherence of one's beliefs, realizing one has made a mistake et cetera.)

The No New Evidence Principle appears to be basic, but on Elga's view it is false. Violating the principle is a high cost to pay since the change in credence on Elga's view is not only strange but runs counter to the project of trying to supply an explanation of how our beliefs should reflect our evidential situation. If Elga is right, then not only is the credo of rationality; "proportion your beliefs to your evidence" violated unwittingly, but in some cases, it ought not to be respected. One receives no new information or evidence and yet one must allocate a probability of one third to the Heads outcome on Monday and one half for the same outcome on Sunday with no change of one's evidential state. It seems, therefore, that the correct characterization of the Sleeping Beauty case is as a

paradox. We find ourselves in the predicament of having to either jettison the No New Evidence Principle or violate the principle that the outcomes of a fair mechanism of chance should be assigned equal credence. Both options are unappealing.

In the present paper our aim is to dispel the paradox by arguing for the one-third answer in a way that makes it clear that there is no problematic violation of the No New Evidence Principle. It has already been noted, by e.g. Arntzenius and Weintraub, that what Elga has to say in regard to the credence shift with no new evidence does not seem satisfactory. (Arntzenius p. 59 and Weintraub, p. 9) His explanation is that "you have gone from a situation in which you count your own temporal location as irrelevant to the truth of [the Heads hypothesis], to one in which you count your own temporal location as relevant to the truth of [the Heads hypothesis]" (Elga, 2000: 145) (He candidly admits, none the less, that one gains no new evidence on his account.)

Ruth Weintraub's account has some affinities with that of the present paper. She maintains that new information *is*, first impressions notwithstanding, gained upon awakening. Namely the information "that one is now awake." Although we will be stating things somewhat different, we will be following her in arguing that information about the waking state is the crucial issue, and furthermore that the ability to pick out Monday indexically *is* important.

A second major claim that we will argue for is that standard Bayesian conditionalization on the hypothesis that "this is an experiment waking-day" (a day on which one is woken up within the Sleeping Beauty experiment) can and should lead one to a credence of one-third in a Heads coin-toss result. We argue for these two major claims in the next two sections and after some concluding remarks we specify in an

appendix how one might deal with the objection that one can demonstratively refer to an experiment waking day on Sunday even though shifting credence in the coin-toss result is not rationally warranted.

Section II: MONDAY EVIDENCE

In order to introduce one of the two main aspects of the present solution to the Sleeping Beauty paradox it is helpful to consider an alternative scenario. In this variant of the Sleeping Beauty experiment there are two coin flips. The first one determines on which day Sleeping Beauty will be woken up, if he is woken up only once. The second coin toss determines whether he will be woken up once or twice. The following matrices display the possible occurrences (Two Coin Flip Scenario):[5]

	Heads1	
	Heads2	Tails2
Monday	W	W
Tuesday	S	W

	Tails1	
	Heads2	Tails2
Monday	S	W
Tuesday	W	W

('S' – Sleeping Beauty will remain sleeping and 'W' he will be woken up.)

In this scenario it is clear that the awakenings will be associated with new information. If sleeping beauty wakes up on Monday, he will know *of* Monday that it is a waking day - even though he will not know *that* Monday is a waking day. And this is something that he could not know beforehand. He has, in other words, new *de re* knowledge about

[5] This is somewhat similar to the red/green light case considered by Weintraub (2004).

Monday; he knows that this day is a waking-day (a day within the experiment on which one will be woken up). This is dependent on his being able to refer to Monday, on his "meeting" Monday as he does in the experiment. Similarly, if he wakes up on Tuesday he will acquire the new information *of* Tuesday that it is a waking day. In this variant case, the problem about new information does not, then, arise. Before flipping the first coin, there was no day that was guaranteed to be a waking day, hence, by being woken up Sleeping Beauty learns of the day in question something that could not have been known beforehand, namely, that this day is an experiment waking-day.

The standard scenario, on the other hand, is associated with the difficulty that Monday *is* known beforehand to be a wakening day. What needs to be explained is how Sleeping Beauty can treat his information, upon awakening on Monday, of Monday that it is a waking-day, as new information. The answer, we submit, is that, for all he knows, it might be genuinely new information, since he does not know that the day is not Tuesday. He is in a situation where he has acquired a certain piece of information - that the present day is an experiment waking-day - but is unable to relate that to another body of information, that body of information that is represented by the calendar. He cannot say which day of the calendar the present day is, and he is therefore, *prima facie*, unable to put these pieces of knowledge together. Our suggestion is that in calculating his credences he should take into account the *probability* of the various possible connections between the different bodies of knowledge. This is a principle that is not dependent on the specifics of the Sleeping Beauty problem, and which is thus of more general applicability.[6]

[6] This is a somewhat simplified formulation of what takes place during the Sleeping Beauty experiment.

Imagine someone challenging Sleeping Beauty on Monday morning claiming that he has gained no new information: "You knew you would be woken up and now you are. What information have you gained?" In reply, he may appeal to the two coin flip case: "Do you agree in that case that I would have gained new information from being woken up?" If the challenger claims that there is no new information in that case either since one knew beforehand that one would be woken up anyhow on one of the two days, Sleeping Beauty can reply that there was no way he could have known that he would be woken up on *this* day: "I could not have known this since you could not have known either. My being woken up today depended on the result of a coin flip. Noticing the result of the coin flip and connecting it with the calendar day you would have learned that I will be woken up today. I only learn this from being woken up, and I still don't know of what day within the experiment I have learned that it is a waking-day. Nevertheless," Sleeping Beauty continues, "I have learned something since if you now would tell me what day it is and nothing more, I would not have to be told the outcome of the two tosses which could not have occurred."[7]

Sleeping Beauty can then go on to argue (as in effect we do), that the original case is only a more complex case of the first: "I do not know that this is not Tuesday, I therefore do not know that I have not gained new coin-flip dependent information. I now no longer know that I do not know of Tuesday that it is a waking day within the experiment." This knowledge is gained by Sleeping Beauty, i.e. knowledge that he no longer knows (as he did know on Sunday), that he does not know of Tuesday that it is a waking-day within the

---

Further refinement will follow as we spell out our account.

[7] For example, if the challenger tells Sleeping Beauty that it is Monday, he would know that the coin did not turn out heads in both coin flips.

experiment.[8] For all he knows it might now be Tuesday (the experiment setup guarantees that one will not know of what day one gains this information).[9]

In the next section we shall describe how Sleeping Beauty should react to the - for him - potentially new knowledge of Monday that it is a waking day.

Section III: CALCULATION OF BEAUTY'S MONDAY CREDENCE

We start with how we think the credence of a Tails toss is to be calculated on Monday, a calculation that resolves the paradox. We will subsequently turn to explaining how this calculation is to be understood and argue for the premises that we think make evident the correctness of the calculation.

In general outline our solution is as follows: On Monday one can demonstratively refer to a day introducing a name for it by means of a demonstrative referring phrase, e.g.. "this day". It is customary to take this name as rigidly designating a day, in this case Monday, even if one does not know what day it is. Moreover, one knows by being awake that the rigidly designated day is an experiment waking-day (a day in the experiment during which one is awake). Calculating the probability of the referred to day being a Monday or a Tuesday provides the correct result.

We define a predicate "*W*(x)" to denote the property of being an experiment waking-day, i.e. "x is a waking-day within the experiment". "*d*" is the name introduced via the demonstrative phrase "this day" (the way one introduces the name can be modified if the

---

[8] There are, of course, many cases where one loses knowledge and yet no credence shift is warranted. What in particular allows for a credence shift in this case is that Sleeping Beauty knows he lost knowledge because he does not know to which day he is referring among the possible two. If he now correctly calculates the probability of referring to each of the possible days as an experiment waking-day, the result will provide the correct posterior credence value.

[9] It might be helpful to put this in terms of the distinction between *de re* and *de dicto* knowledge. For more on this see appendix.

referential access to the definite description changes, e.g. “that day”, “the day we had lunch together”, etc.). “Wd” then, is to be read as: *d is an experiment waking-day*. “Tails” denotes the proposition (or hypothesis) that the coin landed tails (we will focus for the time being on the case where the coin was tossed on Sunday). “Mo” will stand for Monday and “Tu” for Tuesday (this relates to the calendar body of knowledge mentioned in the previous section). Let us now represent the way you can calculate the probability that the coin landed Tails on what we (but not you) know to be Monday. The general idea is to try to calculate the probability that the day you have fixed by a demonstrative phrase is a Tuesday (or Monday), as if asking; “What is the probability of this waking day being a Monday and what is it of being a Tuesday?” The prior probabilities that tie the probability of *d being an experiment waking-day* to a Tails coin toss will deliver the correct answer regarding the credence in the coin toss outcome:

$$\Pr_{new}(Tails)=\Pr(Tails|Wd) \tag{1}$$

$$=\frac{\Pr(Tails)\Pr(Wd\,|\,Tails)}{\Pr(Wd)} \tag{2}$$

$$=\frac{\frac{1}{2}\times 1}{\Pr(Mo=d)\Pr(Wd\,|\,Mo=d)+\Pr(Tu=d)\Pr(Wd\,|\,Tu=d)} \tag{3}$$

$$=\frac{\frac{1}{2}}{(\frac{1}{2}\times 1)+(\frac{1}{2}\times\frac{1}{2})}=\frac{1}{2\times\frac{3}{4}}=\frac{2}{3} \tag{4}$$

Before turning to the question of whether the calculation accurately reflects the reasoning of an ideally rational subject in Sleeping Beauty’s situation, let us set aside some questions of justification. The equation of (1) and (2) is Bayes’ theorem, (3) is just

the conditional probability of *d* being a waking day given Tails (which equals 1 since he will then be woken up on both of the possible days *d* could be). Also, the prior probability of *Tails* is ½ (he knows it's a fair coin). The denominator of (3) is an instance of the rule of total probability: Suppose that you want to establish your credence in an event $A(\alpha)$, where α is a parameter the value of which you do not know with certainty. Suppose, for simplicity, that there are two possible values, 0 and 1. Then $Pr(A(\alpha))= Pr(\alpha=0)Pr(A(\alpha)|\alpha=0)+Pr(\alpha=1)Pr(A(\alpha)|\alpha=1)$. It is made clear in setting up the Sleeping Beauty case that the only possible values for him of *d* are Monday and Tuesday. (3)'s denominator, then, exhausts the possibilities.

With the exception of the justification of the values in (4) - to be given below - our argument is essentially complete. The previous section argues for Sleeping Beauty's having potentially new relevant information, i.e. that this is an experiment waking-day and the present section shows how this information can be used (by standard Bayesian means) to calculate the one-third result. Basically, we have argued that what Sleeping Beauty needs to be asking himself is how probable it is that this day which he knows is an experiment waking-day (by means of his ability to refer and come to know this) is a Monday and how probable it is that it is a Tuesday. In other words he is calculating how probable it is that his knowledge gained by reference relates to his calendar body of information, his knowledge, that is, that these are the only possible values. The calculation shows how one ought to allocate one's credence in such a situation if one is rational. What remains, then, to be explained, in the following section, is the nature of the probability functions involved in the calculation, and the assignment of prior probabilities $Pr(d=Mo)=Pr(d=Tu)= ½$.

Section IV: AN EXPLANATION AND DEFENSE OF THE CALCULATION.

The prior probability we are appealing to in the calculation relates to the time on Monday before taking into account the waking information. To see more clearly what this prior probability amounts to, consider the following variant of the Sleeping Beauty experiment that we will term "BELL":[10]

> The experiment is exactly as the Sleeping Beauty experiment only that in this version you are awake on both days. If a toss of a fair coin results in Tails, a bell will ring on Monday evening and on Tuesday evening at 6:00 and if it will result in Heads, the bell rings at 6:00 pm on Monday but will not ring on Tuesday. You are promised, however, that during the experiment you will not know what day it is. The experimenters will erase your memory when you go to sleep on Monday. As far as you will be able to tell when you wake up on Tuesday, it will be just as if you were waking up on Monday (for the first time within the experiment).

For the purposes of BELL, we exchange the predicate "W" for "B" which will now denote: *x is a bell day* (x is a day on which a bell rings at 6:00 pm as part of the experiment). Now since you are up on Monday (though you don't know it's Monday) you can use the demonstrative indexical phrase "this day" to introduce *d* (which rigidly denotes Monday in the case we are considering) prior to hearing the bell sound. It is easier in BELL to consider your prior probabilities in a way that does not confuse the bell

[10] Thanks to John Hawthorne for bringing this case to our attention.

sound information with the other aspects of the subject's relation to the day. It is now more vivid what your prior probability is for "*d* is a bell day" and you can use it in conditionalizing on *Bd* should you hear the bell sound at 6:00 pm:

$$\Pr_{new}(Tails) = \Pr(Tails \mid Bd) = \frac{\Pr(Bd \mid Tails)\Pr(Tails)}{\Pr(Bd)} \qquad (5)$$

Suppose you do hear a bell. Coming to know that "*d is a bell day*" you can conditionalize as in (1)-(4) to reach a two-thirds Tails probability. The reason things are now somewhat clearer is that the waking/bell information is separated from the demonstrative reference ability in a more conspicuous way. Clearly, you have priors in this case for "*d is bell day*". Encountering a bell sound you conditionalize and shift your credence as a function of the probability that you have encountered the bell sound if *d*=Monday and given *d*=Tuesday. The result of the calculation is the same.

We claim that BELL is analogous to the standard Sleeping Beauty case. Granted, in the latter situation it may be that the priors are not held for very long since the fact of Sleeping Beauty's being awake is quickly, if not immediately, taken into account by him. For those who view this as troubling we specify several ways to alleviate the worry. The priors may be taken to relate to a hypothetical epistemic state preceding the act of factoring in the waking information. But for two reasons hypothetical priors are not a necessary measure for the purposes of accepting our argument. First, as the BELL case illustrates, the connection between the conditionalization and being awake is a superficial feature of the Sleeping Beauty case. The way we are physically constituted could have been different. We just happen to be physically constituted in a way that makes our waking state bound to our reasoning capacities, making us beholden only to credences we have while awake. We could have been constituted in a way that makes the waking signal

a bell sign and while we are asleep we could have been in full command of our rational faculties. Second, even if one views these features as essential for rationality, we could appeal to a different tactic besides hypothetical prior probabilities. Sleeping Beauty can be described as having close to zero credence on Sunday in the proposition: Today is an experiment waking-day. We then can view the waking on Monday as what allows him to conditionalize on this proposition and the calculation of (1)-(4) will remain as before.[11] In any event, it seems like a particularly unhappy position to resist the present account on the grounds that there is no time for the prior probability to be held. Clearly in BELL one can hold a prior probability for the day being a bell day, and claiming that one ought in this case to have a posterior probability of one-third for a Heads coin toss result but not in the original Sleeping Beauty case, seems unreasonable.

A similar point to the one made by an appeal to the BELL case can be made by means of another variant of the Sleeping Beauty story (this case will be useful for other reasons as well). Let us label this version of the experiment "COGNIZER":

> The same experiment as the original Sleeping Beauty experiment is known to you - the cognizer - to be performed on a subject SB. You do not know if the experiment started Sunday night or Saturday night (the time when SB is first put to sleep). You are told Monday morning that SB is now up as part of the experiment but you do not know what the coin toss result is.

[11] We thank a referee for this journal for this suggestion. As this referee pointed out, this way of viewing the prior probability brings out an advantage the present account has over Weintraub's. In her 2004 paper she suggests that the information gained on Monday is represented by the sentence: *I am awake now*. On Sunday, Sleeping Beauty is certain that he is awake (or close to certain) but conditionalizing on this information at that time would hardly generate the same result as it should on Monday. Nevertheless, since the sentence's use (or token thought represented by means of this sentence) expresses different propositions on Sunday and on Monday, there is no real conflict here with Weintraub's account.

Now you can reason just like Sleeping Beauty would. You know that this day, a day you, but not SB, know to be Monday is a waking day. But since you do not know on what day the experiment started, you do not know if the coin toss result of Tails is what is responsible for him now being awake, or not. Your prior for it being a waking day within the experiment was not in any way something you would be rationally advised to conditionalize on. Perhaps you didn't even know that the experiment was taking place until you were told that SB is now awake within the experiment. Or suppose you were only told that SB was asleep within the experiment (you do not know when it began) and shortly after you are told that SB has been woken up. Asked what the probability is of a Tails coin toss result you can appeal to the prior probability that is now unrelated to the waking information.

COGNIZER makes clear another feature of the original Sleeping Beauty paradox that might be responsible for some of the confusion about what can and should be considered as relevant. In the Sleeping Beauty case, one has two functions that in the COGNIZER are separated: First, there is the referring function, the ability under a given protocol[12] to refer to a day as a waking day and demonstratively introduce a name for that day within the timeframe of the experiment. Second, there is the function of calculating the probability (under the given protocol) to be able to refer as one is referring. You as cognizer are also demonstratively referring, via your knowledge that SB is now awake, to a waking day within the experiment. However, since there is no loss of self-location (only the loss of time location of the beginning of the experiment), this case makes the

[12] For more on what we mean by "protocol" see appendix.

existence of the two functions clearer, functions that tend to confuse the issue when not carefully separated. One way to mix them up, and there is a strong tendency to do so, is to confuse the waking information with the day-reference fixing ability. In other words, the ability to refer to a waking day is easily confused with the information or knowledge that one is awake. This is why it might be tempting not to consider the prior probability of Tuesday and Monday as equal. One might think that since I'm awake it is more likely that today is a Monday. Separately considering things, however, makes the equality of these values more evident.

To see why, let us now turn more directly to consider the value distribution given in the denominator of (4). We take it to be uncontroversial that the conditional probability of *d* being a waking day given that it is Monday, i.e. Pr(*Wd*|*d*=*Mo*), is 1. Likewise that the probability of *Wd* given that *d*=Tu, is ½. After all, being a waking day given that it is Monday is stipulated to be certain and it depends on a flip of a fair coin if it's Tuesday.

Doubts, however, may arise when considering the probability of *d* being Monday or Tuesday. The warranted prior probability, we argue, is ½.

One point that must be kept in mind from the outset if mistakes are not to arise is the need to consider the matter without taking in any of the new information about being awake. The probabilities Pr(*d*=*Mo*) and Pr(*d*=*Tu*) are the prior probabilities unrelated to being a waking day, so the waking information should not be factored in. Hence the consideration above that appeals to being awake now is erroneous. But the point that may be raised against our account is that even if it is not factored in, a prior probability of ½ for the unconditional probabilities has not been given sufficient justification.

When we say that the waking information should not be factored in, that means that

there are four relevant possibilities none of which have been eliminated: Heads-Monday, Tails-Monday, Heads-Tuesday, and Tails-Tuesday. Without the waking information a one half prior is warranted.

It might be helpful, as a means of enforcing the plausibility of our priors, to have a look at some ways in which a person might be misled to assign other priors. First, it seems that there may be a temporal partiality at play; the fact that Monday comes before Tuesday (in the experiment), may have something to do with the possible intuition that Monday is more likely than Tuesday. To see why this is a bias, consider the case of Sleeping Beauty but with the days reversed. You will be woken up on Tuesday whether or not the coin lands Heads and you will be woken up on Monday only if the coin lands Tails. If it now seems that Tuesday is more likely than Monday then the waking information is tacitly and illegitimately influencing the judgment. If it still seems that Monday is more likely, then the temporal order is probably behind the intuition (everything else is set to be equal). But it ought not have this effect.[13]

---

[13] We can construct Sleeping Beauty like cases where the temporal order of events will not, we think, have the same intuitive effect since order of time will be in opposition to what event is "closer". For instance, the experimenters tell you that you will remember on Wednesday only one of the awakenings but you will not know which. Now, the closer event in time – a possible or actual Tuesday waking - is the last event in ordinal time from the Wednesday perspective. So now two aspects can be separated from each other: the closer event and the order of events. Now, say you vividly remember one of the awakenings. Abstracting away the fact that your awake on that day, what do you think the probability is of it being a Monday? (If the waking information gets in the way think of the BELL case from this Wednesday perspective.)

The Wednesday example brings out another interesting feature of the Sleeping Beauty case. This perspective (from Wednesday) is exactly like your situation on Sunday if you are promised and in fact do not remember any of the wakings and it is like your Monday/Tuesday perspective if you are promised that you will only remember one awakening and that the memory will be random. You can then demonstratively introduce a name for the remembered day and calculate your new credence in accordance with (1)-(4). If you are promised to remember only the last waking (be it Monday or Tuesday) your credence should be half for a Heads/Tails toss. This is due to your reference protocol changing: depending on a flip of a fair coin toss, *d*=Monday or *d*=Tuesday.

It has been pointed out to us by a referee for this journal that a one half prior credence for it being Tuesday (and for *d*=Monday) is justified by the indifference principle that both Elga and Lewis accept. See footnote 4 above. The idea is that if the same propositions are true in both cases (or the same non-centered possibilities are actual), one will be warranted in dividing one's credence between the two centered

Second, let us again look at the BELL variant of the Sleeping Beauty experiment. In that case, one does not know whether it is Monday or Tuesday and one knows that one will be awake on both days. The only relevant difference between the days is that in one day, that is Monday, the bell will ring in any event at 6:00 pm, while on Tuesday it will ring only if the coin-toss results in Tails. Intuitively, when the other information is abstracted away it seems rational to give a one half prior credence to the day's being Tuesday. If you don't think so, reverse the order of the bell-coin dependence such that the bell will ring on Monday (not Tuesday) depending on a Tails result (just like in the previous case we considered). It seems that when the waking/ringing information is taken away and the days are reversed, a prior of ½ is the best answer.

We have argued, first, that the result of the drug's making the day on which one is awake inaccessible from a first person perspective is to randomize the reference to a particular day. Under the protocol that is stipulated for the Sleeping Beauty case, the correct prior probability upon awaking of a day being a Monday or a Tuesday should be one half. We do not, however, claim that this is mandated for all cases since one might have some independent justification for a different prior probability that will justify another value for the coin toss result. The probability that one gets as a result of conditionalization in the manner we have advocated will be as justified as one's prior probability is regarding what day it is (the other values seem to be fixed). But there is no such justification in the Sleeping Beauty case and hence we have claimed that when the other information that might distract from the correct prior is removed, the one half prior appears to be the only justified value for this case. In any event the result that Lewis was

---

possibilities. What this means is that Lewisians cannot question the prior probability of *d*=Tuesday. For a further reason why Lewisians are in trouble in this respect, see the main text bellow.

after seems entirely unwarranted since it demands absolute certainty that *d*=Mo and other answers will be warranted only if one shifts the reference protocol, i.e. the known conditions under which one gets to refer. This means that in order to get a posterior probability for the coin toss result of one half using the calculation above, the prior probability of *d*=Monday needs to be 1. Whether you agree or not with our claim that the prior probability of *d*=Monday should be one half, the value Lewisians need with respect to the calculation is entirely unwarranted.

Second, we have argued that there is no real problem with regard to when the prior probability is held and how. There are different ways we specified of how one might view this question and we do not think it necessary in this context to settle the issue of which one is best. For instance, we have adopted the view that one can have a low probability (that does not warrant conditionalization) on Sunday for: *Today is an experiment waking day*, which then upon awakening will be suited for conditionalization.

Section V: CONCLUSION

We briefly recap in these concluding remarks and bring our argument's two main claims into clearer overall view. We have argued in section II, mainly by appealing to the two coin flip scenario, that given the ability Sleeping Beauty has, upon waking up but no earlier, to demonstratively refer to a waking day within the experiment, he in fact gains potentially new relevant information. To be more exact, as we further specify in an appendix to this paper, it is not the mere ability to demonstratively refer that is of informative content since one does, or can, have demonstrative reference in cases where a shift is not warranted. Rather, it is the ability to demonstratively refer under a *certain*

protocol that can, and as we have argued in this case, does, provide relevant new information that can be utilized to warrant a credence shift. So besides arguing that in the Sleeping Beauty case one potentially gains information, we also explain how that happens, and why it can be used for conditionalization. This claim of a rationally warranted credence shift as a result of potentially new information is the one resisted by both Lewis and Elga. It is hard to see, however, how one might reject the claim that in two-coin flip scenario there is new information and even harder to see how once this is granted one could still claim that in the Sleeping Beauty case there is no change in the epistemic situation that warrants the credence shift. Moreover, it seems straightforward that "this day is an experiment waking-day" is informative and once it is made clear (as we hope to have done in the paper and in the appendix) how it is informative, we see no reason why one would not be able to conditionalize on this proposition.

The second major claim of our argument is that the conditionalizing will give a one-third warranted credence for a Heads coin-toss result. This result, which coincides with Elga's one-third answer, is reached by standard Bayesian means in the third section. We also defended this result in face of two initial worries. First, that there is no clear sense in which one can have prior credence in the proposition we utilized. Second, that the values we appeal to regarding the prior credence of *d*=Monday, are unwarranted. In response to these worries we have utilized two further variants of the Sleeping Beauty experiment – BELL and COGNIZER – to show that, first, there is no real worry about the prior credence (since there is no apparent problem of prior credence in the BELL case). Second, we have shown that when we neutralize the distracting features in the Sleeping Beauty case, that is, the time order, the waking information, and proximity, the only

value that seems warranted for a *d*=Monday prior, is one half.

If these two main claims work, Sleeping Beauty ought to have credence of one-third that the coin flip result is Heads given the information that he has on Monday that "today is a waking day within the experiment."

Appendix:

Something more detailed can be said about the epistemic state that allows the calculation of credences in the Sleeping Beauty case. In this appendix we will be addressing a worry concerning one's situation regarding the future credence from a Sunday perspective. To do so we will first introduce in different terms, which might be instructive, the major claims we have made in this paper. This will help us address the worry and further elaborate on notions that we have appealed to in the main text.

The description of the type of knowledge that reflects the conditionalization on Monday is intricate. It involves a certain sort of higher order knowledge of one's ignorance as to what day one is referring to. But even though you are ignorant about what day it is that you are referring to, you are not ignorant as to the possibilities of what you are referring to, and given a distribution of prior probability to those possibilities, you can calculate (as we have specified above) the probability of the coin toss result.

As mentioned in the main text, this is a special case of a more general principle regarding the relation between different bodies of knowledge, where the epistemic subject does not have access to the exact nature of the relevant connections, but can nevertheless jointly employ the different bodies of knowledge in calculating his

credences by means of using the probabilities that he associates with the various possible connections (assuming he knows that they exhaust the sample space).

A way in which this can be represented is by means of the well-known *de dicto* and *de re* distinction. Your knowing that *d* is a waking day will entail *de re* knowledge *of* Monday (if *d*=Monday) that it is a waking day. This knowledge does not derive from a principle that allows one to derive *de re* from *de dicto* knowledge, but rather, from the ability to demonstratively introduce a name *d* for a day that you do not know to be a Monday or a Tuesday. On Monday you do not know of Tuesday that it is a waking day, but you do know that you don't know that you don't know of Tuesday that it is a waking day. Let us explain why, by starting with your knowledge on Tuesday (should you be awake then).

On Tuesday you don't know that you don't know of Tuesday that it is a waking day. The reason you don't know this is simply because you do know of Tuesday that it's a waking day. In other words, since knowledge is factive you can't know that you don't have *de re* Tuesday knowledge of Tuesday that it's a waking day. Nevertheless, since you know you don't know either on Monday or on Tuesday that today is not Tuesday, you can reason and come to know (both on Monday and on Tuesday) that you don't know that you don't know of Tuesday that it is a waking day. That is, both on Tuesday and Monday you have third order knowledge of your second order ignorance of your first order ignorance:

$K_{Mo \vee Tu}$: You know that you don't know that you don't know of Tuesday that it is a waking day.

And this knowledge of second order ignorance stems from your knowledge of two

things: one, that you know that *d* is a waking day, and two, that you don't know of which day you have this knowledge. So in these terms, since you can calculate how probable it is that your *de re* knowledge is of the Monday and Tuesday in the Sleeping Beauty situation, you can get the required result we argued for above.

But if this is all correct, why can't one make this calculation already on Sunday? After all, you know on Sunday that you will be able to refer to Monday since you will be woken up that day no matter what. You know that you will be able to refer to a waking day then, and since the calculation reflects this predicament that you know you will find yourself in, you can update your credence in a Tails coin-toss *now*.

One tempting reply to this quandary would be to claim that since on Sunday you know that you can't refer *de re* to a Tuesday waking day your credence shift would not be warranted. But this reply although it might be adequate for some cases is not one that you can utilize in general.[14]

Let us imagine that the coin has been tossed Sunday morning though you don't know what the toss's outcome is. Now you introduce *l* via the stipulation: "Let *l* refer to the last waking day". Now on Sunday, you know of *l* that it is a waking day, and you know that you don't know that you don't know of Tuesday that it is a waking day. Moreover, you know that the only possible values for *l* are Monday or Tuesday. So it seems that all the conditions we have been appealing to are in place and a credence shift to one-third is now (on Sunday) rationally warranted. But it is absurd to shift credence just because the toss of the coin has taken place, hence our argument must be wrong.

The problem with this challenge is not that it does not warrant the calculation but

[14] Many thanks here to Ofra Magidor for raising a related issue.

rather that the calculation does not give a value that is different from the one you already have for the coin toss result (from a Sunday perspective). The protocol – the way in which you get to refer – is such that a credence shift is not warranted. To see this let us get back to the conditionalizing schema:

$$\Pr_{new}(Tails) = \Pr(Tails \mid Wl) \quad \text{(i)}$$

$$= \frac{\Pr(Tails)\Pr(Wl \mid Tails)}{\Pr(Wl)} \quad \text{(ii)}$$

$$= \frac{1/2 \times 1}{\Pr(Mo = l)\Pr(Wl \mid Mo = l) + \Pr(Tu = l)\Pr(Wl \mid Tu = l)} \quad \text{(iii)}$$

Up until this point, the calculation is the same and is warranted by the same features that we have been employing. However, since *l* was introduced as a last *waking day*, there is no way that it could have failed to be a waking day.[15] Hence the values in the denominator will be as follows:

$$= \frac{1/2}{(1/2 \times 1) + (1/2 \times 1)} = \frac{1/2}{1} = 1/2 \quad \text{(iv)}$$

So even though one could conditionalize on the knowledge that "*d* is an experiment waking-day", the access one has to this event does not warrant a credence shift. In other words, the reference protocol, the way in which the reference to a waking day is fixed, does not warrant a credence shift.

We have seen that a credence shift by way of directly referring to a waking day from a

[15] As pointed out to us by a referee for this journal, another way to put this is that the evidence is old evidence, i.e. by definition *l* is a waking day and hence Pr(W*l*)=1. Conditionalizing on (W*l*) is hence pointless.

Sunday perspective holds no promise. Yet one might think there are other ways - van Frassen's reflection principle (1985)[16] – that would warrant a Sunday credence shift if our argument is correct. Would a reflection principle warrant a Sunday credence shift?

For one thing you don't yet have the ability to refer to *d* demonstratively. But as we saw this is not quite the essential obstacle for a warranted Sunday credence shift. The problem is not that you can't refer demonstratively *per se*, but rather that the way you come to refer on Monday to Monday is by means of a protocol[17] that determines a credence shift. On Sunday if you do have means of reference to a Monday waking day,

---

[16] Van Frassen's reflection principle states (at least in close approximation that is sufficient for the purposes of the current question) that if you are certain that your rational credence will be x in the future, you ought rationally to shift your credence now: $Pr(A|Pr_t(A)=x)=x$ (where "$Pr_t(\bullet)$" is your rational credence at a future time t).

[17] In general what we mean by *protocol* is that there are ways of gaining the ability to demonstratively refer that differ from others - at times warranting different credences. One ought, we maintain, treat one's ability to refer (if one knows how probable it is that one would be able to refer as one actually can) as a measuring devise akin to other measuring instruments that give uncertain information. Two questions may be raised in relation to how we have employed this idea in our account of the Sleeping Beauty paradox. First, is reference relevant for conditionalization in other cases or is it just a special feature of the sleeping beauty case? Second, how do different reference protocols influence credences?

In response to these questions, there is a variety of cases that would illustrate the generality of the notions we have been appealing to under the heading of *demonstrative reference protocol*. One example has already been given in this appendix but here is another example that is not directly related to the Sleeping Beauty case: Consider the well-known case first introduced in Gardner (1959). You know that Smith has two children at least one of them you know to be a boy. What is the probability that Smith has two boys? The answer is one third (assuming there is no other relevant background information). There are three equally probable cases: Smith has a younger boy and an older girl, a younger girl and older boy, or two boys. Second case: You meet Smith with his child that he introduces as his son. What now is the probability that Smith has two boys? The answer here depends on the protocol of demonstrative reference. If you have an equal chance of meeting Smith's other child regardless of the child's sex, the correct answer is one half. The other child is equally likely to be a boy or a girl. But if, say, Smith would only be seen with a son, e.g. Smith adheres to a biased custom according to which "respectable" fathers will not be seen with a girl in public, then you have no chance of meeting him with a girl and should not change your credence to one half (assuming you know about this terrible custom). You ought, under this protocol of reference, to stick to your original one-third credence for Smith's having two sons. You treat your ability to refer as a measuring tool calculating the probability that you would meet one of Smith's children (male or female). Many other examples would do equally well (e.g. the two aces problem). Reference protocol, then, is a general probabilistic phenomenon that is not special to the sleeping Beauty paradox. If one is careful about the protocol of reference one can see how it systematically changes the credences in a way that reflects the points we have been making with regard to this paradox.

For a more comprehensive discussion of the examples see, Maya Bar-Hillel and Ruma Falk R. (1982). And for an argument for the indispensability of direct reference in solving some probability puzzles, see Martine Nida-Rümelin (1992).

e.g. by introducing a name for the first waking day, this reference comes together with extra knowledge that blocks the conditionalization from providing a rational credence shift. The protocol of reference to a waking day on Sunday does not allow you to update your credence since you know on Sunday that you are not referring by *d* (introduced by the method above) to a Tuesday. Hence the calculation in the denominator would not reflect your epistemic state on Sunday (plugging values into Bayes' theorem makes this point even more vivid). In other words your ability to rigidly refer to a waking day needs to come with knowledge of your second order ignorance of what day it is that you are referring to in order to allow the (1)-(4) calculation to reflect your new warranted epistemic state. This is why $K_{Mo \vee Tu}$ is important.[18]

References:

Arntzenius, F. [2002]: "Reflections on Sleeping Beauty." Analysis, 62: 53–62.

Elga, A. [2000]: "Self-locating Belief and the Sleeping Beauty Problem." Analysis, 60: 143–147.

Elga A. [2004]: "Defeating Dr. Evil with Self-Locating Belief." Philosophy and Phenomenological Research Vol. LXIX, No. 2: 383-396.

Bar-Hillel M. and R. Ruma Falk [1982]: "Some Teasers Concerning Conditional Probabilities." Cognition, 11: 109-122.

Gardner, M. [1959]: *Scientific American Book of Mathematical Puzzles and Diversions*. New York, Simon and Schuster, 49-51.

Lewis, D. [2001]: "Sleeping Beauty: Reply to Elga." Analysis, 61: 171–176.

[18] Many thanks to Frank Arntzenius, Adam Elga, David Enoch, John Hawthorne, Ofra Magidor, Peter Pagin, Ariel Rubenstein, Assaf Sharon, Ruth Weintraub and Jonas Åkerman for helpful comments, suggestions and objections that greatly improved the argument of this paper. We jointly presented a previous version of this paper at the Stockholm University Philosophy of Science Seminar – we wish to thank the organizers and participants specifically Per-Erik Malmnäs, Dugald Murdoch and Paul Needham. Special thanks to two referees for this journal who made us think much harder about Sleeping Beauty.

Nida-Rümelin, M. [1993]: “Probability and Direct Reference.” Erkenntnis, 39: 51-78.

Piccione, M. and A. Rubenstein [1997]: “On the Interpretation of Decision Problems with Imperfect Recall.” Games and Economic Behavior, 20: 3–24.

van Fraassen, B. [1984]: “Belief and the will.” Journal of Philosophy, 81: 235–256.

Weintraub, R. [2004]: “Sleeping Beauty: a Simple Solution.” Analysis, 64: 8–10.