Stalnaker on Sleeping Beauty

Brian Weatherson

The Sleeping Beauty puzzle provides a nice illustration of the approach to self-locating belief defended by Robert Stalnaker in *Our Knowledge of the Internal World* (Stalnaker, 2008), as well as a test of the utility of that method. The setup of the Sleeping Beauty puzzle is by now fairly familiar. On Sunday Sleeping Beauty is told the rules of the game, and a (known to be) fair coin is flipped. On Monday, Sleeping Beauty is woken, and then put back to sleep. If, and only if, the coin landed tails, she is woken again on Tuesday after having her memory of the Monday awakening erased.[1] On Wednesday she is woken again and the game ends. There are a few questions we can ask about Beauty's attitudes as the game progresses. We'd like to know what her credence that the coin landed heads should be

(a) Before she goes to sleep Sunday;
(b) When she wakes on Monday;
(c) When she wakes on Tuesday; and
(d) When she wakes on Wednesday?

Standard treatments of the Sleeping Beauty puzzle ignore (d), run together (b) and (c) into one (somewhat ill-formed) question, and then divide theorists into 'halfers' or 'thirders' depending on how they answer it. Following Stalnaker, I'm going to focus on (b) here, though I'll have a little to say about (c) and (d) as well. I'll be following orthodoxy in taking $\frac{1}{2}$ to be the clear answer to (a), and in taking the correct answers to (b) and (c) to be independent of how the coin lands, though I'll briefly question that assumption at the end.

An answer to these four questions should respect two different kinds of constraints. The answer for day *n* should make sense 'statically'. It should be a sensible answer to the question of what Beauty should do given what information she then has. And the answer should make sense 'dynamically'. It should be a sensible answer to the question of how Beauty should have updated her credences from some earlier day, given rational credences on the earlier day.

As has been fairly clear since the discussion of the problem in Elga (2000), Sleeping Beauty is puzzling because static and dynamic considerations appear to push in different directions. The static considerations apparently favour a $\frac{1}{3}$ answer to (b). When Beauty wakes, there are three options available to her: It is Monday and the coin landed heads; It is Monday and the coin landed tails; It is Tuesday and the coin landed tails. If we can argue that each of those are equally probable given her evidence, we get the answer $\frac{1}{3}$. The dynamic considerations apparently favour a $\frac{1}{2}$

[†] Penultimate draft only. Please cite published version if possible. Final version published in *Philosophical Studies* 155 (2011): 445-456. Thanks to Adam Elga, Elizabeth Harman, Ishani Maitra, Ted Sider, Robert Stanlaker and Seth Yelcin for comments on an earlier, and mistake-riddled, draft.

[1] Note that I'm not assuming that Beauty's memories are erased in other cases. This makes the particular version of the case I'm discussing a little different to the version popularised in Elga (2000). This shouldn't make any difference to most analyses of the puzzle, but it helps to clarify some issues.

answer to (b). The right answer to (a) is $\frac{1}{2}$. Nothing happens on Monday or Tuesday that surprises Beauty. And credences should only change if we are surprised. So the right answer to (b) is $\frac{1}{2}$.

Since we must have harmony between dynamic and static considerations, one of these arguments must be misguided. (In fact, I think both are, to some degree.) These days there is a cottage industry of 'thirders' developing accounts of credal dynamics that accord with the $\frac{1}{3}$ answer to (b).[2] But all of these accounts are considerably more complex than the traditional, conditionalisation-based, dynamic theory that we all grew up with.

Three of the many attractions of Robert Stalnaker's new account of self-locating knowledge are (i) that it offers a way to answer all four of our questions about Sleeping Beauty, (ii) that it does so while remaining both statically and dynamically plausible, and (iii) that the dynamic theory involved is, in large part, traditional conditionalisation. I spend most of this note setting out Stalnaker's account, and setting out his derivation of a $\frac{1}{3}$ answer to (b). I conclude with some reasons for preferring a slightly different solution of the Sleeping Beauty puzzle within the broad framework Stalnaker suggests.

1 Stalnaker on Self-Location

The picture of self-locating belief that we get from Lewis's "Attitudes *De Dicto* and *De Se*" (Lewis, 1979) has been widely adopted in recent years.[3] On Lewis's picture, the content of an attitude is a set of centered worlds. For current purposes we'll take to centered worlds to be $\langle$world, agent, time$\rangle$ triples. To believe that *S*, where *S* is a set of centered worlds, is to believe that the triple $\langle$your world, you, now$\rangle$ $\in$ *S*.

The motivation for this picture comes from reflection on how to represent locational uncertainty. If you're sure where in New York City you are, you can pick out a point on a map and say "I'm there". If you're not sure exactly where you are, but you have some information, you can pick out a region on the map and say "I'm somewhere in that region". If you're not sure who you are, but you know where everyone is, you can do the same kind of thing. And it's plausible that this is a (somewhat) realistic situation. As one modern-day Lewisian, Andy Egan, puts it 'I can believe that my pants are on fire without believing that Egan's pants are on fire, and I can hope that someone turns a fire extinguisher on me right now without hoping that someone turns a fire extinguisher on Egan at 5:41pm." (Egan, 2004, 64) There is an important puzzle here that needs to be addressed, and can't obviously be addressed in the framework Lewis accepted before 1979, where the content of a propositional attitude is a set of Lewisian concreta. If possible worlds are Lewisian concreta, then Lewisians like Egan are correct to respond to puzzles about location by saying, "sometimes (as when we want to know who or where we are) **the world is not enough**". (Egan, 2004, 64)

But this response is too self-centered. Not all locational thoughts are self-locational thoughts. I can be just as uncertain about where *that* is as about where *this*

[2] See, for instance, Titlebaum (2008) and the references therein.

[3] Including by me. See Egan et al. (2005)

is, or as uncertain about who *you* are as about who *I* am. Imagine I'm watching Egan's unfortunate adventures with his infernal trousers on a delayed video tape. I can believe *his* pants are on fire without believing Egan's pants are on fire, and hope that someone turns a fire extinguisher on him *then* without hoping some turns a fire extinguisher on Egan at 5:41pm. Or, at least, that way of putting things sounds just as good as Egan's original description of the case.

For a different example, imagine I wake at night and come to believe it is midnight. As Lewis would represent it, I believe $\langle w, \text{me}, \text{now}\rangle \in \{\langle w, s, t\rangle : t = \text{midnight}\}$. When I wake, I think back to that belief, and judge that I may have been mistaken. How should we represent this? Not that I now believe $\langle w, \text{me}, \text{now}\rangle \notin \{\langle w, s, t\rangle : t = \text{midnight}\}$. That's obviously true - I know the sun is up. We want to represent something more contentious.

The best, I think, the Lewisian can do is to pick out some description D of my earlier belief and say what I believe is $\langle w, \text{me}, \text{now}\rangle \notin \{\langle w, s, t\rangle : (\iota x : Dx)x \text{ happens at midnight}\}$. That is, I believe the belief that satisfies D doesn't happen at midnight. Is that good enough? Well, we might imagine the debate continuing with the anti-Lewisian proposing cases where D will not be unique (because of forgotten similar beliefs) or will not be satisfied (because of a misrecollection of the circumstances of the belief), and so this approach will fail. And we might imagine the Lewisian responding by complicating D, or by denying that in these cases we really do have beliefs about our earlier beliefs. In other words, we can imagine the familiar debates about descriptivism about names being replayed as debates about descriptivism about prior beliefs. As enjoyable as that may be, it's interesting to consider a different approach.

There's a more philosophical reason to worry about Lewis's model. If we model uncertainty as a class of relationships to possible worlds, it looks like there's a lot of actual uncertainty we won't be able to model. Indeed, there are three kinds of uncertainty that we can't model in this framework. First, we can't model uncertainty about logic and mathematics. Second, if we accept the necessity of identity, we can't model uncertainty about identity claims. Whatever it is to be uncertain about whether a is b, it won't be a distinctive relation to the set of worlds in which a is b, since that's all the worlds. Third, we can't model uncertainty about claims about self-identity, like *I'm that guy*. Lewis's framework is an improvement on the sets of possible worlds approach because it helps with this third class of cases. But it doesn't help with the first or, more importantly, with the second. We might think that a solution to puzzles about self-identity should generalise to solve puzzles about identity more broadly. Lewis's model doesn't. One of Stalnaker's key insights is that we should, and can, have a model that addresses both kinds of puzzles about identity.

On Stalnaker's model, a belief is just a distinction between worlds. The content of a belief is a set of worlds, not a set of centered worlds. But worlds have more structure than we thought they had. The formal model is a bit more subtle than what I'll sketch here, but I think I'll include enough detail to cover the Sleeping Beauty case. In each world, each center, in Lewis's sense, has a haecceity. A world is the Cartesian product of a Lewisian world, i.e. a world without haecceities, and a function from each contextually salient haecceity to a location. If we see a kiss, and wonder who *she* is, who *he* is, and *when* they are kissing, then we can think of the worlds as quadruples consisting of a haecceity-free world (perhaps a Lewisian concreta), a woman, a man and a time. So we can represent three kinds of locational doubts, not just self-locational doubt.[4]

When an agent at center *c* believes something self-locating, e.g. that it is Monday, the content of their belief is that *c*'s haecceity is on a Monday. If they don't know what day it is, there's a sense in which they don't know what they believe, since they don't know whether what they are believing is that *c*'s center is on Monday, or that some other center's haecceity is on Monday.[5] But their belief, the belief they would express on Monday by saying "It is Monday", has two nice features. First, it is neither trivial, like the belief that *Monday is Monday*, nor changing in value over time, since *c*'s center is always on Monday. Second, it is the kind of belief that people on days other than Monday can share, or dispute. And this belief can be shared by others who have the capacity to think *de re* about *c*, even if they can't uniquely describe it. It's this last fact that lets Stalnaker handle the cases that proved problematic for Lewis and the neo-Lewisians. For instance, it lets Stalnaker model shared uncertainty about identity claims.

With all that in place, it's time to return to Sleeping Beauty. Let's consider two propositions. The first, *H*, is that the coin landed heads. The second, *M*, is what Beauty can express when she wakes on Monday by saying "It is Monday". That is, it is a singular proposition about a wakening experience that Beauty can now have singular thoughts about (since she is now undergoing it), but which she didn't previously have the capacity to determinately pick out. We'll call this wakening *a*. (Beauty might undergo multiple wakenings, but we're going to focus on one for now, and call it *a*.) Given these three propositions, we can describe four possibilities. Or, as we'll somewhat inaccurately describe them, four worlds.[6]

w_1: $H \wedge M$
w_2: $H \wedge \neg M$
w_3: $\neg H \wedge M$
w_4: $\neg H \wedge \neg M$

On Sunday, Beauty's credences are distributed over the algebra generated by the partition $\{H, \neg H\}$, i.e., $\{\{w_1, w_2\}, \{w_3, w_4\}\}$. The algebra is that course-grained because she doesn't have the capacity to think *M* thoughts. And that's because she's not acquainted with the relevant haecceities. So she can't distinguish between worlds that differ only on whether *M* is true. On Sunday then, Beauty's credences are given by $Pr(H) = Pr(\neg H) = \frac{1}{2}$.

[4] Stalnaker thinks we have independent reason to treat these structured entities as simply worlds. The main point of the last few sentences was that we can adopt Stalnaker's model while staying neutral on this metaphysical question.

[5] Perhaps it would be better to say that individuals and times have haecceities, rather than saying centers do. I have little idea what could tell between these options, or even if there is a substantive issue here.

[6] Of course worlds are considerably more detailed than this, but the extra detail is an unnecessary confusion for the current storyline.

When she wakes on Monday, two things happen. First, she becomes acquainted with a. So she can now think about whether a is on Monday. That is, she can now think about whether M is true. So she can now carve the possibility space more finely. Indeed, now her credences can be distributed over all propositions built out of the four possibilities noted above. The second thing that happens is that Beauty rules out one of these possibilities. In particular, she now knows that $H \wedge \neg M$, a proposition she couldn't so much as think before, is actually false. That's because if the coin landed heads, this very wakening could not have taken place on Tuesday.

Stalnaker's position on Beauty's credences uses these two facts. First Beauty 'recalibrates' her credences to the new algebra, then she updates by conditionalising on $\neg H \vee M$. If after recalibration, her credences are equally distributed over the four cells of the partition, the conditionalising on $\neg H \vee M$ will move $Pr(H)$ to $\frac{1}{3}$. That is, the thirders win!

But we might wonder why we use just this calibration, the one where all four cells get equal credence. We're going to come back to this question below. But first, I want to use Stalnaker's framework to respond to an interesting objection to the thirder position.

2 Monty Hall

Both C. S. Jenkins (2005) and Joseph Halpern (2004) have argued that the 'thirder' solution is undermined by its similarity to fallacious reasoning in the Monty Hall case. The idea is easy enough to understand if we simply recall the Monty Hall problem. The agent is in one of three states s_1, s_2 or s_3, and has reason to believe each is equally likely. She guesses which one she is in. An experimenter then selects a state that is neither the state she is in, nor the state she guessed, and tells her that she is not in that state. If she simply conditionalises on the content of the experimenter's report, then her credence that she guessed correctly will go from $\frac{1}{3}$ to $\frac{1}{2}$. This is a bizarre failure of Reflection, so something must have gone wrong.[7] Both Jenkins and Halpern suggest that the violation of Reflection that 'thirders' endorse in Sleeping Beauty is just as bizarre.

But the Sleeping Beauty puzzle is not analogous to the Monty Hall problem. That's because in Sleeping Beauty we seem forced to have a violation of Reflection somewhere. Let's think a bit again about Beauty's credences on Wednesday, and let's assume that we're trying to avoid Reflection violations. Then on Monday (and Tuesday) her credence in H is $\frac{1}{2}$. Now when Beauty awakes on those days, there are three possibilities open to her. (Hopefully it won't lead to ambiguity if I re-use the name a for the awakening Beauty is undergoing when thinking about H.)

- a is Monday and H
- a is Monday and $\neg H$
- a is Tuesday and $\neg H$

[7] The standard response is to say that the agent shouldn't just conditionalise on the content of the experimenter's utterance, but on the fact that the experimenter is making just that utterance. We'll return to this idea below.

When she wakes on Wednesday, she's in a position to reflect on these possibilities. And she can rule out the second of them. That's what she learns when she wakes and learns it is Wednesday; that if $\neg H$, then that last awakening was on Tuesday. Now since that last awakening, nothing odd has happened to Beauty. She hasn't had her memories erased. She might have had her memories erased between Monday and Tuesday, but that's not relevant to the time period she's considering. Moreover, she knows that she hasn't had her memories erased. So I think she's in a position to simply conditionalise on her new evidence. And that new evidence is simply that whatever else was going on when she was thinking about those three possibilities, she wasn't in the second possibility.

But now we face a challenge. Beauty knows that Wednesday will come. So if her credence in H on Wednesday isn't $\frac{1}{2}$, then we'll have a violation of Reflection. The violation is that on Sunday her credence in H is $\frac{1}{2}$, but she knows it will go up on Wednesday. And that violation is just as bad as the violation of Reflection that 'thirders' endorse. But if she conditionalises when she wakes up on Wednesday, then the only way her updated credence in H can be $\frac{1}{2}$ is if her prior credence in the first and third options above were equal. And the only way that can happen is for her credence, when a is happening, in the proposition that a is Monday and $\neg H$ is 0. But that's bizarre. Whether or not the thirders are right to think that she should give equal credence to that possibility as to the two others, she can't give it credence 0. So Reflection will fail somewhere.

To see why Reflection is failing in these cases, it helps to look back at the requirements we need in order to get from conditionalisation to Reflection. In Rachael Briggs's careful analysis of when Reflection holds, in Briggs (2009), Reflection is only guaranteed to hold when agents know what their evidence is. In other cases, even perfect conditionalisers may violate Reflection.

This assumption, namely that agents know what their evidence is, is a kind of luminosity assumption. And not surprisingly, it has been challenged by Timothy Williamson (Williamson, 2000, 230-3). What is a little more surprising is that we only need a relatively weak failure of luminosity in order to get problems for reflection. The assumption that agents know what their evidence is can be broken into two parts.

- If p is part of S's evidence, then S knows that p is part of her evidence.
- If p is not part of S's evidence, then S knows that p is not part of her evidence.

The first part is, I think, implausible for reasons familiar from Williamson's work. But the second is implausible even if one doesn't like Williamson's style of reasoning. If we think p must be true to be part of S's evidence (as I think we should), and we think that rational agent's can have false beliefs about anything, as also seems plausible by simple observation of how easy it is to be misled, then even a rational agent can fail to realise that p is not part of her evidence. The easiest way that can happen is if she falsely, but reasonably, believes p, and hence does not realise that due to its falsity, it is not part of her evidence.

Williamson provides an interesting model, based on a discussion in Shin (1989), of a case where an agent does not know that something is not part of her evidence.

There are currently three possible states the agent could be in: s_1, s_2 or s_3. An experiment will be run, and after the experiment the agent will get some evidence depending on which state she's in.

- If she's in s_1, her evidence will rule out s_3.
- If she's in s_2, her evidence will rule out s_1 and s_3.
- If she's in s_3, her evidence will rule out s_1.

Assume the agent knows these conditionals before the experiment is run, and now let's assume the experiment has been run. Let xRy mean that y is possible given the evidence S gets in x. Then we can see that R is transitive. That means that if p is part of S's evidence, then her evidence settles that p is part of her evidence. But R is not Euclidean. So it is possible that p is not part of her evidence, even though her evidence does not settle that p is not part of her evidence. In particular, if she is in s_1, that she isn't in s_1 is not part of her evidence. But for all she can tell, she's in s_2. And if she's in s_2, her evidence does rule out her being in s_1. So her evidence doesn't settle that this is not part of her evidence.

The model is obviously an abstraction from any kind of real-world case. But as we argued above, it is plausible that there are cases where an agent doesn't know what evidence she *lacks*. And this kind of case makes for Reflection failure. Assume that the agent's prior credences are (and should be) that each state is equally likely. And assume the agent conditionalises on the evidence she gets. Then her credence that she's in s_2 will go up no matter what state she's in. And she knows in advance this will happen. But there's no obvious irrationality here; it's not at all clear what kind of reflection-friendly credal dynamics would be preferably to updating by conditionalisation.[8]

So when an agent doesn't know what evidence she lacks, Reflection can fail. One way to think about the Sleeping Beauty case is that something like this is going on, although it isn't quite analogous to the Shin-Williamson example discussed above. In that example, the agent doesn't know what evidence she lacks at the *later* time. In the Sleeping Beauty case, we can reasonably model Beauty as knowing exactly what her evidence is when she wakes up. Her evidence does nothing more or less than rule out w_2. That's something she didn't know before waking up. But in a good sense she didn't know that she didn't know that. That's because she was not in a position to even think about w_2 as such. Since she wasn't in a position to think about a, couldn't distinguish, even in thought, between w_1 and w_2. So any proposition she could think about, and investigate whether she knew or not, had to include either both w_1 and w_2, or include neither of them. So the only way she could know that

[8] The idea that we should update by conditionalisation on our evidence, even when we don't know what the evidence is, has an amusing consequence in the Monty Hall problem. The agent guesses that she's in s_i, and comes to know she's not in s_j, where $i \neq j$. If she only comes to know that she's not in s_j, and not something stronger, such as knowing that she knows she's not in s_j, then she really should conditionalise on this, and her credence that her guess was correct will go up. This is the 'mistaken' response to the puzzle that is frequently deprecated in the literature. But since the orthodox solutions to the puzzle rely on the agent reflecting on how she came to know $\neg s_j$, it seems that it is the right solution if she doesn't know that she knows $\neg s_j$.

she didn't know $\{w_1, w_3, w_4\}$ is if she tacitly knew she didn't know that in virtue of knowing that she didn't know $\{w_1, w_2, w_3, w_4\}$. But she didn't know that she didn't know that for the simple reason that she did know that $\{w_1, w_2, w_3, w_4\}$, i.e. the universal proposition, is true. So we have a case where Beauty doesn't know what it is she doesn't know at the earlier time. And like cases where the agent doesn't know what she doesn't know at the later time, this is a case where reflection fails.

So there are two reasons to be sceptical of reflection-based arguments against the 'thirder' solution to the Sleeping Beauty puzzle.

- There is no plausible way for Beauty's credence in H to be $\frac{1}{2}$ on both Monday and Wednesday, but reflection requires this.
- Reflection is only plausible when agents know both what evidence they have, and what evidence they lack, throughout the story. And it is implausible that Beauty satisfies this constraint, since she gains conceptual capacities during the story.

But this isn't a positive argument for the $\frac{1}{3}$ solution. I'll conclude with a discussion of two arguments for the $\frac{1}{3}$ solution. Both arguments are suggested by Stalnaker's framework, but only one of them is ultimately defensible.

3 Stalnaker on Sleeping Ugly

When we left Stalnaker's discussion of the Sleeping Beauty case, we had just noticed that there was a question about why Beauty should respond to being able to more finely discriminate between states by 'recalibrating' to a credal state where each of w_1 through w_4 receive equal credence. This question about calibration is crucial to the Sleeping Beauty puzzle because there are other post-calibration distributions of credence are are *prima facie* viable. Perhaps, given what Beauty knows about the setup, she should never have assigned any credence to $H \wedge \neg M$. Rather, she should have made it so $Pr(\neg H \wedge M) = Pr(\neg H \wedge \neg M) = \frac{1}{4}$, and $Pr(H \wedge M) = \frac{1}{2}$. If she does that, the conditionalising on $\neg(H \wedge \neg M)$ won't change a thing, and $Pr(H)$ will still be $\frac{1}{2}$. That is, the halfers win!

One argument against this, and in favour of the equally weighted calibration, is suggested by Stalnaker's 'Sleeping Ugly' example. Sleeping Ugly is woken up on Monday and again (with erased memories) on Tuesday however the coin lands. So when Ugly awakes, he has the capacity to think new singular thoughts, but he doesn't get much evidence about them. In particular, he can't share the knowledge Beauty would express by saying, "If the coin landed Heads, this is Monday."[9] Now we might think it is intuitive that Ugly's credences when he wakes up and reflects on his situation should be equal over the four possibilities. Moreover, *all* Ugly does is recalibrate; since he doesn't learn anything about which day it is, his post-awakening credence just is his recalibration. If all this is correct, and if Beauty should recalibrate in the same way as Ugly, then Beauty should recalibrate to the 'equally weighted calibration'. And now we're back to victory for the thirders!

[9] Stalnaker notes that this is a reason for thinking Beauty does learn something when she wakes up, and so there's a reason her credence in H changes.

But there's little reason to believe the crucial premise about how Ugly should recalibrate his credences. What we know is that Ugly doesn't have any reason to give any more credence to any one of the four possibilities than to the others. It doesn't at all follow that he has reason to give equal credence to each, any more than in general an absence of reasons to treat one of the *X*s differently to the others is a reason to treat them all the same.[10]

The argument I'm considering here is similar to reasoning Adam Elga has employed Elga (2004), and which I have criticised Weatherson (2005). A central focus of my criticism was that this kind of reasoning has a tendency to lead to countable additivity violations. In an important recent paper, Jacob Ross (2010) has shown that many thirder arguments similarly lead to countable additivity violations. He shows this by deriving what he calls the 'Generalised Thirder Principle' (hereafter, GTP) from the premises of these arguments. The GTP is a principle concerning a generalised version of the Sleeping Beauty problem. Here is Ross's description of this class of problems.

> Let us define a *Sleeping Beauty problem* as a problem in which a fully rational agent, Beauty, will undergo one or more mutually indistinguishable awakenings, and in which the number of awakenings she will undergo is determined by the outcome of a random process. Let *S* be a partition of alternative hypotheses concerning the outcome of this random process. Beauty knows the objective chances of each hypothesis in *S*, and she also knows how many time she will awaken conditional on each of these hypotheses, but she has no other relevant information. The problem is to determine how her credence should be divided among the hypotheses in *S* when she first awakens. (Ross ms, 2-3)

The GTP is a principle about this general class of problem. Here's how Ross states it.

Generalized Thirder Principle In any standard Sleeping Beauty problem, upon first awakening, Beauty's credence in any given hypothesis in *S* must be proportional to the product of the hypothesis' objective chance and the number of times Beauty will awaken conditional on this hypothesis. ... [We can] express this principle formally. For any hypothesis $i \in S$, let $Ch(i)$ be the objective chance that hypothesis i is true, and let $N(i)$ be the number of times Beauty awakens if i is true. Let P be the Beauty's credence function upon first awakening. The GTP states ...

$$\text{For all } i, j \in S, \frac{P(i)}{P(j)} = \frac{N(i)Ch(i)}{N(j)Ch(j)} \text{ whenever } Ch(j) > 0. \text{ (Ross ms, 6-7)}$$

[10] Compare this argument for giving nothing to charity. There are thousands of worthwhile charities, and I have no reason to give more to one than any of the others. But I can't afford to give large equal amounts to each, and if I gave small equal amounts to each, the administrative costs would mean my donation has no effect. So I should treat each equally, and the only sensible practical way to do this is to give none to each. Note that you really don't have to think one charity is more worthy than the others to think this is a bad argument; sometimes we just have to make arbitrary choices.

The argument I'm considering seems to be committed to the GTP. In a generalised Sleeping Beauty problem, we can imagine a version of Sleeping Ugly who will awake every day that Beauty might awake. The reasoning that leads one to think that Ugly should give equal credence to each of the two days in the original Sleeping Beauty case seems to generalise to imply that Ugly should give equal credence to each day in this more general case. But if in the general example Beauty calibrates to match these credences of Ugly, then conditionalises on the information she receives, then she'll end up endorsing the GTP. That's an unhappy outcome. It would be better to have an argument for the $\frac{1}{3}$ solution that doesn't imply the GTP.

I'm going to argue that when Beauty wakes up her credences should satisfy the following two premises. (As always, I use *a* to name the awakening that Beauty is now undergoing, and I'm using *Cr* for her credence function on waking.)

P1: $Cr(a \text{ is Monday and } H) = Cr(a \text{ is Tuesday and } \neg H)$
P2: $Cr(a \text{ is Monday and } H) = Cr(a \text{ is Monday and } \neg H)$

These constraints imply, given what Beauty knows about the setup, that $Cr(H) = \frac{1}{3}$. The arguments for each premise are quite different.

The argument for P1 is one I mentioned above, so I'll just sketch it quickly here. On Wednesday, Beauty's credence in *H* should be back to $\frac{1}{2}$. But what she learns on Wednesday is $\neg(a \text{ is on Monday and } \neg H)$. So on Monday, her credence in *H* conditional on $\neg(a \text{ is on Monday and } \neg H)$ should be $\frac{1}{2}$. But given what Beauty knows about the setup of the problem, this immediately implies P1.

The argument for P2 requires a slightly more fanciful version of the example. Imagine that on Sunday night, Beauty is visited by a time traveller from Monday who comes back with a videotape of her waking on Monday, and tells her that it was taken on Monday. So Beauty now has the capacity to think about this very awakening, i.e., *a*. This doesn't seem to affect her credences in *H*, it should still be $\frac{1}{2}$. Now imagine that her memory of this visit is erased overnight, so when she wakes up on Monday her situation is just like in the original Sleeping Beauty problem.

Call Cr_1 her credence function on Sunday after meeting the time traveller. And call Cr_2 her credence function on Monday after she wakes up and reflects on her situation. It seems the only relevant difference between the situation on Sunday and the situation on Monday is that Beauty has *lost* the information that *a* is on Monday. The following principle about situations where an agent loses information seems plausible. If Cr_{old} is the pre-loss credence function, and Cr_{new} is the post-loss credence function, and *E* is the information lost, then

- $Cr_{\text{old}}(p) = Cr_{\text{new}}(p|E)$

The idea here is that information loss is a sort of reverse conditionalisation. Applying this, we get that $Cr_1(H) = Cr_2(H|a \text{ is Monday})$, so $Cr_2((H|a \text{ is Monday}) = \frac{1}{2}$, so $Cr_2(a \text{ is Monday and } H) = Cr_2(a \text{ is Monday and } \neg H)$. And since the situation on Monday in the revised problem, i.e., the situation when Beauty's credence function is Cr_2 is just like the situation in the original Sleeping Beauty problem on Monday, it follows that P1 is true in the original problem. And from P1 and P2, it follows that the thirder solution is right.

But note a limitation of this solution. When Beauty wakes on *Tuesday* her credence function is defined over a different algebra of propositions to what it was defined over after meeting the time traveller. So there's no time travel based argument that her credences on Tuesday should satisfy P2, or indeed that on Tuesday her credence in H should be $\frac{1}{3}$. (For similar reasons, this kind of reason does not support the GTP.)

One might try and argue that Beauty's situation on Tuesday is indistinguishable from her situation on Monday, and so she should have the same credences on Tuesday. Both the premise and the inference here seem dubious. On Tuesday, Beauty knows different singular propositions, so the situation isn't clearly indistinguishable. But more importantly, it is implausible that indistinguishability implies same credences. The relation *should have the same credences in* is a transitive and symmetric relation between states. The relation *is indistinguishable from* is neither transitive nor symmetric. So I suspect that the kind of arguments developed here leave it an open question what Beauty's credences should be on Tuesday, and indeed whether there is a unique value for what her credences then should be.

References

Briggs, Rachael, (2009). “Distorted Reflection.” *Philosophical Review* 118: 59-85, doi:10.1215/00318108-2008-029. (6)

Egan, Andy, (2004). “Second-Order Predication and the Metaphysics of Properties.” *Australasian Journal of Philosophy* 82: 48-66, doi:10.1080/713659803. (2)

Egan, Andy, Hawthorne, John, and Weatherson, & Brian, (2005). “Epistemic Modals in Context.” In Gerhard Preyer and Georg Peter (eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, 131-170. Oxford: Oxford University Press. (2)

Elga, Adam, (2000). “Self-Locating Belief and the Sleeping Beauty Problem.” *Analysis* 60: 143-147, doi:10.1093/analys/60.2.143. (1)

—, (2004). “Defeating Dr. Evil with Self-Locating Belief.” *Philosophy and Phenomenological Research* 69: 383-396, doi:10.1111/j.1933-1592.2004.tb00400.x. (9)

Halpern, Joseph, (2004). “Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems.” In *Oxford Studies in Epistemology*, volume 1, 111-142. Oxford: Oxford University Press. (5)

Jenkins, C. S., (2005). “Sleeping Beauty: A Wake-Up Call.” *Philosophica Mathematica* 13: 194-201, doi:10.1093/philmat/nki015. (5)

Lewis, David, (1979). “Attitudes *De Dicto* and *De Se*.” *Philosophical Review* 88: 513-543, doi:10.2307/2184646. Reprinted in *Philosophical Papers*, Volume I, pp. 133-156. (2)

Ross, Jacob, (2010). "Sleeping Beauty, Countable Additivity, and Rational Dilemmas." *Philosophical Review* 119: 411-447, doi:10.1215/00318108-2010-010. (9)

Shin, Hyun Song (1989). "Non-partitional Information on Dynamic State Spaces and the Possibility of Speculation." Working Paper 90-11, Center for Research on Economic and Social Theory, Univesity of Michigan. (6)

Stalnaker, Robert, (2008). *Our Knowledge of the Internal World*. Oxford: Oxford University Press. (1)

Titlebaum, Michael, (2008). "The Relevance of Self-Locating Beliefs." *Philosophical Review* 117: 555-605, doi:10.1215/00318108-2008-016. (2)

Weatherson, Brian, (2005). "Can We Do Without Pragmatic Encroachment?" *Philosophical Perspectives* 19: 417-443, doi:10.1111/j.1520-8583.2005.00068.x. (9)

Williamson, Timothy, (2000). *Knowledge and its Limits*. Oxford University Press. (6)