# Sleeping Beauty's Evidence

Jeffrey Sanford Russell

April 2020

Here's the familiar story of Sleeping Beauty (Elga 2000, 143):

> Some researchers are going to put you to sleep. During the two days that your sleep will last [Monday and Tuesday], 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?

I'll start from the idea that the degree to which Beauty *ought to believe* the coin came up Heads is the same as the degree of belief which is *supported by Beauty's evidence.* This hardly sounds innovative, but it encourages a little reorientation. Most of the literature on this puzzle focuses on *update rules* that relate Beauty's beliefs on Sunday to her beliefs at later times. Thinking in these terms makes questions about propositions that are only sometimes true (or only sometimes entertainable) seem very important (see Titelbaum 2013 and references therein). If we focus instead on the import of the evidence Beauty has *on Monday,* and what this evidence supports, things get easier in some ways (compare the "time-slice" approach of Moss 2012; Hedden 2015).

Here are some lessons I'll discuss. First, part of Beauty's evidence on Monday is that she is awake; we can argue straightforwardly that this evidence is relevant to the outcome of the coin flip, given other things she knows. Second, Beauty's evidence can provide one-sided confirmation for Tails, without any possibility of confirmation for Heads. Third, Beauty might have *additional* evidence that screens off the evidence that she is awake from the coin flip; whether this is so depends on contentious issues about the nature of evidence, transparency, and defeat.

I'll use "you *can tell* that *P*" as a convenient shorthand for "you have evidence that entails *P*."

Thanks to Charity Anderson, Andrew Bacon, John Hawthorne, Alex Pruss, a seminar at the University of Southern California, and an audience at Baylor University for helpful discussion. Special thanks to Bernhard Salow for detailed comments.

## 1 Relevant Evidence

Let's warm up with a variant story. You wake up disoriented, completely unsure where you are or what day it is. Of course you do notice

**Woke.** You woke up today.

You find a stack of three index cards. (You can tell they are trustworthy.) You begin to read them, one by one.

**Day.** It is either Monday or Tuesday.
**Coin.** A fair coin has been flipped.

At this point, the probabilities given your evidence look like Table 1.

Table 1: Probabilities supported by Woke, Day, and Coin

| | Monday | Tuesday |
|---|---|---|
| Heads | 1/4 | 1/4 |
| Tails | 1/4 | 1/4 |

This much seems obvious, but it's worth pausing to ask why. I *don't* think it is because rationality requires that you satisfy some general indifference principle that applies in this case. Such principles face well-known difficulties (for example van Fraassen 1989). Rather, I think Table 1 is correct because *this is an artificial probability puzzle.* We have conventions for filling in details in vignettes like these. If the story says a goat is behind one of three doors, we understand each door to be equally likely on your evidence (as specified so far in the story). We also understand which door it is to be independent of questions like which door you decide to open. Life doesn't *have* to be this way. You could have inside information, or Monty Hall might look suspicious, or you might suspect that your choice is subliminally guided by a faint goatish smell. But those would be perverse interpretations of the story, as a puzzle. We might call these guiding conventions *hermeneutic indifference principles.* But it would be a mistake to conclude from the conventions of puzzle stories that we always *are* in situations where our background probabilities for arbitrary questions we don't know much about are uniform and independent. When it comes to applying lessons from puzzles to real life cases, we will have to reconsider whether these assumptions are appropriate.[1]

We now turn to the third index card.

**Protocol.** You do not wake up on Tuesday if the coin lands Tails.

Equivalently:[2]

> It is not the case that (the coin came up Heads and it's Tuesday and you woke up today).

The import of Protocol given your other evidence is easy to work out. It rules out the possibility Heads-and-Tuesday, while telling you nothing new about any of the other three cases. If we conditionalize Table 1 on Protocol, we get Table 2. So Woke, Day, Coin, and Protocol together support Heads to degree 1/3.

Table 2: Probabilities supported by Woke, Day, Coin, and Protocol

| | Monday | Tuesday |
|---|---|---|
| Heads | 1/3 | 0 |
| Tails | 1/3 | 1/3 |

This reasoning relied on the principle that we can work out what some evidence supports by breaking it up into pieces.

**Combined Evidence.** For propositions $E$ and $H$, let $P_E(H)$ be the degree of belief in $H$ supported by evidence $E$. Then

$$P_{E_1 \& E_2}(H) = P_{E_1}(H \mid E_2)$$

[1] Here are two reasons for belaboring this point.

This explanation for Table 1 does not depend on Elga's "highly restricted principle of indifference," that qualitatively indiscernible predicaments within a single possible world must receive equal probability (Elga 2000, 144).

Hermeneutic indifference principles can come into conflict. For example, if a story introduces a natural countable partition, the convention "assume uniformity" conflicts with the convention "assume coherence" (including countable additivity). I take this to be what is going on in the variant in Ross (2010), which involves infinitely many days (see also Weatherson 2011). Ross argues that the case involves a "credal dilemma" where different *epistemic* norms are in conflict (supposing a principle like Elga's). I conclude instead that different *interpretive* norms for the story conflict, and so it's just not clear what the probabilities are in the story without more to go on.

[2] I stipulate that the "if" on the Protocol card is a material conditional. More on this in Footnote 9.

(as long as $P_{E_1}(E_2) \neq 0$).

We first worked out the probabilities supported by the evidence *without* Protocol (Table 1), and then using Combined Evidence we can deduce the probabilities supported by all of the evidence together.[3]

Combined Evidence is not a *diachronic* principle. We imagined reading the index cards one by one, but this was just to help make certain conditional probabilities vivid. In particular, while our reasoning did involve conditionalizing one probability function to calculate another, it does not rely on the principle of *(diachronic) conditionalization.* That would be tendentious in this context (see Titelbaum 2013, sec. 4).

Combined Evidence fits better with this alternative picture (compare "Synchronic Conditionalization" in Hedden 2015; see Meacham 2016 and references therein):

**Ur-Prior.** There is a probability function $P_\top$—call it the *ur-prior*—such that, for any propositions $H$ and $E$,

$$P_E(H) = P_\top(H \mid E)$$

(as long as $P_\top(E) \neq 0$).

The idea is that the ur-prior encodes once and for all the relations of evidential support that different propositions bear to one another. Many have been skeptical that there is any such thing (for example, Ramsey [1926] 2010). I'm inclined to treat such worries not as showing that there is no ur-prior, but rather as suggesting that different probability functions might play the "ur-prior role" for different people, or at different times, or that it might be vague or otherwise hard to tell which probability function plays the ur-prior role. (Keynes (1921) and Carnap (1950) thought of the ur-prior as "logical" in some important sense, but that isn't part of the Ur-Prior picture I am discussing.)

Combined Evidence implies Ur-Prior, by considering the case of tautologous evidence $\top$.[4] But it might be sensible to restrict Combined Evidence. Maybe it doesn't make any sense to have only tautologous evidence—excluding even contingent evidence like *I exist* or *I have some evidence*. There are various restrictions of Combined

[3] Two background assumptions are baked into the notation. The first is that the relevant evidence consists of *propositions*. The second is that what is supported by some evidence propositions taken together is the same as what is supported by the *conjunction* of those propositions.

[4] The converse almost holds; using the standard ratio definition of conditional probability, Ur-Prior implies Combined Evidence for cases where $P_\top(E_1) \neq 0$.

Evidence to "reasonable" evidence propositions $E_1$ and $E_2$ that would suffice for the work we're putting it to. So while it's a natural fit, we don't need to rely on the full-fledged Ur-Prior idea.

Another thing to notice is that since all our reasoning was about evidence at one time, it doesn't matter that your evidence has a special temporal profile. Maybe the evidence "I woke up today" is a "centered proposition" (as Elga took for granted), or maybe it is a proposition which is only sometimes true (see for example Sullivan 2014), or maybe that sentence expresses different propositions at different times, and so corresponds to different evidence on different days (for example Weatherson 2011). This makes no difference to our reasoning. Since we are only paying attention to what your evidence *now* supports, it could work any of these different ways equally well.

Now back to Sleeping Beauty. Is her situation relevantly different from yours in the variant story? Upon waking on Monday, she has all of the evidence you have in the story: she can tell that she woke up today, that it is Monday or Tuesday, that a coin was to be flipped, and that she only wakes on Tuesday if the coin comes up Tails. So parallel reasoning tells us that *this much* of her evidence supports degree of belief one-third in Heads. It is true that she gained this evidence in a different way than you did—there were no index cards, and instead she was told Protocol on Sunday, rather than learning it after she woke on Monday. However she got it, though, Beauty's evidence on Monday includes the conjunction of Woke, Day, Coin, and Protocol.

Beauty also has some *additional* evidence that didn't come into the variant story. She didn't wake up entirely disoriented, and instead she knows she is involved in an experiment that could involve memory erasing. This might make a difference—I'll consider one way it might matter in Section 2. So we haven't settled the question in favor of the one-third answer. But we have gained some ground.

Consider a standard objection to the Thirder:

> Between Sunday night and Monday morning, all she learns is 'Today is Monday or Tuesday.' Information about the passage of time doesn't seem relevant to facts about coin flips, so Beauty's credence in heads on Monday morning shouldn't change from what it was on Sunday night − namely 1/2. (Titelbaum 2013, 1006; compare Lewis 2001)

But this is not all she learns. She also learns that she woke up today; given the protocol, this information is relevant to Heads (compare Weintraub 2004).[5]

[5] Elga (2000, 145) stipulatively uses the word "information" in a restricted sense: "an agent receives

One way to make the relevance more vivid is to imagine a bystander with Beauty who also knows the experimental protocol, who is awake both days no matter what, but who also can't tell whether it is Monday or Tuesday (for whatever reason—maybe they also get their memory erased Monday night; compare Stalnaker 2008, 63). When the bystander observes that Beauty wakes up, they thereby straightforwardly gain evidence against Heads. (Given Heads she might not have woken up then, whereas given Tails she certainly would.) But when Beauty wakes up, she also observes this, and thus gains access to the very same fact that the bystander does, with the same evidential import.

There is an important difference between Beauty and the bystander. If Beauty had *not* woken up, the bystander would have gained this evidence instead—that Beauty didn't wake up—and thereby gained evidence *for* Heads. But Beauty would not have gained this alternative evidence in that case. This brings us to a second standard objection to the Thirder. Beauty has an opportunity to gain evidence *against* Heads, but no opportunity to gain evidence *for* Heads. That conflicts with this principle:

**Confirmation Balances Out.** If you could gain evidence against $H$, then you could gain evidence for $H$.

This seems troubling. One way to press the point is to imagine Beauty thinking through the reasoning on Sunday evening: then she assigns Heads probability one-half, but she can anticipate that whatever happens, her probability for Heads upon waking on Monday will be one-third. This is an odd mismatch—a violation of the *reflection principle*. Why not just skip ahead and adopt credence one third on Sunday?

The reply is that Confirmation Balances Out is not generally true. Here's a simpler case to illustrate this point. If I'm popular, I want to gain evidence for this. If I'm not, I'd rather remain ignorant. You know whether I'm popular. I can't just *ask* you—then you might tell me I'm unpopular, which would make me sad. But I can put you up to this plan: if I'm popular, wake me up tonight; otherwise, let me sleep. If I find myself awake, I will have gained evidence for my popularity. But there is no danger of my finding myself asleep![6] (Alas, my happy epistemic state will be

new information when she learns the truth of a proposition expressible by an eternal sentence of some appropriately rich language." Titelbaum does not use the word in this restricted way—he allows here that there is such a thing as "information about the passage of time"—and neither do I.

[6] This example is inspired by Salow (2018), who argues that such "intentionally biased inquiry" is impossible. Note that Salow rules out inquiry that involves *forgetting*—and he might argue (with Shakespeare's Henry IV) that being asleep will "steep my senses in forgetfulness." I don't know about this, but the example illustrates the present point about imbalanced confirmation whether or not it is a true counterexample to Salow's claim. (The firing squad from Leslie 1989 has a similar structure, though it is usually deployed for different purposes.)

short-lived. If I *ever* wake up and remember that you didn't wake me, evidence for my unpopularity will come home to roost. Eternal slumber is a high price to pay for permanent one-sided confirmation.)

The basic asymmetry, at the heart of the Sleeping Beauty puzzle, is that when you're awake, you normally can tell that you're awake, but when you're asleep, you normally can't tell that you're asleep. Indeed, if (like the bystander) Beauty *could* tell whenever she was asleep, things would not be very puzzling (compare the variants in Arntzenius 2003, sec. 4; Dorr 2002; Weintraub 2004). In that case, on Tuesday Beauty might get the evidence that she didn't wake up, which would raise the probability of Heads to one (given the protocol). This possible confirmation would balance out the possible disconfirmation in the other three cases. But the fact that Beauty would sleep through her only shot at getting evidence that *confirms* Heads doesn't mean she also loses her chance to get evidence that *disconfirms* Heads when she is awake on Monday.

There is a theorem of the probability calculus in the vicinity of Confirmation Balances Out:

**Confirmation Theorem.** If $\mathscr{E}$ is a set of mutually exclusive and jointly exhaustive propositions (a *partition*), and some $E \in \mathscr{E}$ disconfirms $H$ (where $E$ has positive probability), then some $E \in \mathscr{E}$ confirms $H$.[7]

But the principle that Confirmation Balances Out only follows from the Confirmation Theorem if the set of propositions that you could have as evidence is a partition of the possibilities compatible with your current evidence. And this is not always so. (Compare the closely related discussion in Williamson 2002; Briggs 2009; Weatherson 2011.)

In the simple "fishing for compliments" case, I might get evidence that tells me I'm awake and popular; but while I might well be asleep and unpopular, in this case I won't get any evidence that tells me so. So my possible bodies of evidence do not partition my possible predicaments. In the "good case" where I'm awake, my evidence rules out being asleep, while in the "bad case" where I'm asleep, I do not have evidence that rules out being awake.

The Sleeping Beauty case is more complicated, because it also involves losing track of time, but the essential feature is the same as the simple case: the propositions that

[7] *Proof.* Recall that $P(H)$ is equal to a weighted average of the conditional probabilities $P(H \mid E)$ for $E \in \mathscr{E}$ such that $P(E) \neq 0$. So if one of these conditional probabilities is lower than $P(H)$, some other conditional probability must be higher than $P(H)$.

might turn out to be Beauty's total evidence do not partition her possible predicaments. In the "good cases" where she is awake (Monday & Heads, Tuesday & Heads, or Tuesday & Tails) she has evidence that rules out the "bad case" where she is asleep (Tuesday & Heads). But in that "bad case," she will not have evidence that rules out being awake. The conditions of the Confirmation Theorem do not apply, and it is false that Confirmation Balances Out. The Confirmation Theorem tells us that if you can get evidence against Heads, then there is some alternative proposition that would count in favor of Heads—but it does not follow that this alternative proposition is also one that you might ever have an opportunity to *learn*.[8]

The usual challenge for the Thirder is to explain how any of Beauty's evidence is relevant to the coin flip. Now the tables are turned. Beauty does have *some* evidence conditional on which the probability of Heads is 1/3. So defenders of *non-Third* answers face a challenge: what *extra* evidence besides Woke, Day, Coin, and Protocol might she have that shifts the probability away from 1/3 to something else?[9]

## 2 Wide Evidence

According to one view, **Evidence is Narrow**: your evidence is entirely determined by features of your present qualitative phenomenal state—the way things appear to you. Thus your evidence is something you can share with (for example) a brain in a vat which is stimulated in just the right way. An alternative view says that **Evidence**

[8] The complications of time require care. Since we are now explicitly comparing evidence at different times, we have to attend to the nature of temporary evidence. Suppose on Sunday Beauty has evidence *it's Sunday* which is incompatible with her later evidence *it's Monday or Tuesday*—so this is a case of evidence loss. Then the evidence she might have on Monday does not partition the possibilities compatible with her Sunday evidence for the trivial reason that this evidence is *incompatible* with her Sunday evidence. This issue would equally arise even in perfectly ordinary cases when time passes. By itself, it need not utterly block the Confirmation Theorem: we can also consider confirmation with respect to intermediate hypothetical bodies of evidence. Consider the evidence that consists of Beauty's Sunday evidence *minus* the evidence that it's Sunday, and *plus* the evidence that it's Monday or Tuesday—but without the additional evidence that Beauty is awake. Beauty never *has* just this evidence, but we can still consider abstractly what it would support. But with respect to *this* evidence, we still cannot apply the Confirmation Theorem, for the reason just discussed.

[9] Bernhard Salow has suggested (in personal correspondence) that it is natural to interpret the protocol conditional ("You don't wake up on Tuesday if the coin lands Heads") as stronger than a mere material conditional. The additional modal information it conveys—something like "*the protocol ensures* that you don't wake up on Tuesday when the coin lands heads"—might in principle be relevant to the coin flip. It is possible to construct a prior that includes various "protocol propositions" like this such that, without the protocol proposition as evidence, the probabilities of the four cells of the table are uniform, but adding the protocol proposition as evidence supports halfing. I can't give this idea the attention it deserves here; my own judgment is that priors with this structure are not well-motivated, but this is an interesting route for the halfer to pursue.

**is Wide.** In a slogan, evidence ain't in the head. ("Internalism" and "externalism" are common labels for these views.) In particular, your evidence includes facts about the world that you observe and remember:

**Observation is Evidence.** If you observe that *P*, then it's part of your evidence that *P*.

**Memory is Evidence.** If you remember that *P*, then it's part of your evidence that *P*.

(The way I've characterized them, Evidence is Narrow and Evidence is Wide aren't exhaustive alternatives: evidence might be wider than your phenomenal state, but not so wide as to include all that you observe or remember. But they are mutually exclusive, because whether you observe that *P* is not determined by your present phenomenal state. I observe that I'm holding a pen, but my phenomenal duplicate in a vat does not observe that they are holding a pen, and instead merely *seems* to observe this.)

One well-known wide evidence view is the "E=K" thesis: "knowledge, and only knowledge, constitutes evidence" (Williamson 2002, 185). What you know is not a matter of your present phenomenal state. But the thesis that Evidence is Wide can come apart from E=K. First, you may know things that aren't evidence. For instance, it's natural to think that knowledge which you have *inferred* from observations shouldn't be "double-counted" as evidence itself. Second, you may have evidence beyond what you know. Williamson (2002 ch. 1) argues that if you observe that *P*, or remember that *P*, then you know that *P*; but I'm not sure whether this is true. Sometimes seeing people don't believe their eyes. Such a person might perfectly well *observe* that it's raining, and thereby have as part of their *evidence* that it's raining—but even so they don't *believe* that it's raining, and thus they don't know it. If this can happen, and Observation is Evidence, then some evidence is not knowledge. (Similar remarks apply to memory.)[10] So the wideness of evidence is not tied specifically to E=K.

Wide evidence provides resources for a distinctive response to skepticism (see Williamson 2002, ch. 8). Norma has strong evidence that she is a university professor. For example, she observes that she teaches students, submits articles to journals, attends committee meetings, and so on. Skip is a phenomenal duplicate

[10] Here's an alternative diagnosis. Maybe sometimes you *don't* see that *P*, but you *can* see that *P*. (Compare Williamson's notion of what you're *in a position to know*.) We might think your evidence doesn't just include what you *do* see, but also what you *can* see. (We can similarly distinguish what you remember from what you *can* remember.)

of Norma subject to a *Truman Show* style hoax: her classroom is full of actors, the articles she writes don't go to journals, and the committees whose meetings she attends are entirely fictitious. If Evidence is Wide, then Norma has evidence that Skip doesn't, and the attitudes that are supported by Norma's evidence are different from those supported by Skip's. It's compatible with Skip's evidence that she has no students—indeed, she doesn't! Skip *can't tell* whether she has students; but Norma *can* tell that she has students, just by looking around her classroom.

(Note that I'm assuming that the contents of observation and memory are themselves "wide," or "thick," in another sense. You sometimes observe that a student asked a question, or remember that you defended your dissertation; you don't *merely* observe or remember that beings who looked a certain way made certain noises.[11])

Let's call this reply to the skeptic **Asymmetric Dogmatism.** Normal evidence (the *good case*) *rules out* being in a skeptical scenario (the *bad case*). But in a skeptical scenario, your evidence does not rule out either the good case or the bad case.

Sleeping Beauty's situation is also a kind of skeptical scenario. Let's examine how these ideas apply.

For a warm-up, think about when you woke up this morning. How did you know what day it was? In my case, I remembered that yesterday was Sunday (and that yesterday I did my usual Sunday things); I concluded that today was Monday.

Of course, I could have been in alternative predicaments: covert agents from the Philosophical Defense Force could have drugged me to sleep through an entire day, or they could have wiped my memory of an entire day. My usual Monday-morning phenomenology did not by itself rule out these skeptical hypotheses. But this morning when I woke up I wasn't in any of those devious scenarios. It was a normal case, not a skeptical one. If Evidence is Wide, then I had normal "good case" evidence, such as that yesterday was Sunday.

(Again, I take it that the contents of memory are "thick": what you remember isn't just a slideshow of qualitative imagery. I remember that I broke my arm when I was six, that I worked on this draft last Friday—and I remember that yesterday was Sunday.)

On Monday morning, Sleeping Beauty's cognitive story so far is very much like mine. It was a perfectly normal night: she was briefed, she went to sleep, and she woke up. It is plausible that she *remembers* that she was briefed on the experiment yesterday. She had a good night's sleep and woke up refreshed and ready for more

[11] Compare the "cognitive penetrability of experience" (Siegel 2012, *inter alia*)—though that discussion focuses on the content of "narrow" perceptual experience itself, rather than "wide" states like observation and memory.

epistemic science. It seems that all her cognitive processes are in perfectly good order to deliver her the evidence that yesterday was Sunday, and thus that today is Monday. If *my* evidence implies that today is Monday, then why wouldn't Sleeping Beauty's as well?

Of course, sometimes you can't tell what time it is. Waking up in the trunk of a car, you may have no idea how long you've slept, or how often you've woken (Elga 2004). If Beauty wakes on *Tuesday*, after a night of devious neurological tampering, she wakes with a mind full of temporal illusion, and thus plausibly with much less evidence than upon an ordinary waking. Does the mere *threat* of such tampering in the future suffice to undermine her memories on Monday? This is not clear. Tuesday wakings and Monday wakings are not symmetric. Tuesday is a skeptical scenario in relation to Beauty's normal waking on Monday. If Evidence is Wide, it may well be that on Tuesday Beauty can't tell what day it is, and on Monday she can—even if her experiences on the two days are qualitatively identical. If Evidence is Wide, the question of whether Beauty can tell what day it is *on Monday* is open.

Some people will think that Beauty's ordinary waking memories on Monday are *defeated* by other information she has about the experimental protocol (compare Elga 2004; Arntzenius 2003, 356). There are three ways this defeat could work. (1) Beauty does have evidence that entails that it is Monday, but she should not have the degrees of belief that her evidence supports. I set this option aside at the outset: what we are presently exploring are the attitudes that are supported by Sleeping Beauty's *evidence.* (2) Beauty remembers that yesterday was Sunday, but it is not part of her evidence that yesterday was Sunday. This option conflicts with the thesis that Memory is Evidence, so I set it aside as well. The remaining option: (3) Beauty does not remember that yesterday is Sunday. Call this option *Memory Defeat.* She experiences an *apparent* memory that yesterday is Sunday—just as she will if she wakes on Tuesday—but unlike that case, this experience is perfectly veridical, and produced by the ordinary cognitive processes that produce genuine memories, without any "cognitive mishap." Nonetheless, the view is that this is not enough to qualify it as a genuine memory. Cognitively, nothing unusual disrupted Beauty's peaceful slumber on Sunday night. (Direct memory tampering only happens Monday night in this experiment, if at all.) Epistemically, though, it is as if someone had erased her memories, and then replaced them with apparent memories that by accident had the very same content as the originals.

It is tempting to look for a fourth option: Beauty remembers that yesterday was Sunday, and does thereby have this as part of her evidence, but additionally she has *further* evidence which lowers the probability again. But that's a non-starter. If *some* of Beauty's evidence raises the probability that it is Monday all the way to *one*, then no further evidence can lower it again.[12] One way to argue for this appeals again to the Combined Evidence principle. If $P_{E_1}(\text{Monday}) = 1$, then for any further evidence $E_2$

$$P_{E_1\&E_2} = P_{E_1}(\text{Monday} \mid E_2) = 1$$

(as long as $P_{E_1}(E_2) \neq 0$). If Beauty's memories (on their own) merely made it highly *probable* that it was Monday, then they could be defeated on their own terms by further probability-lowering evidence. But if Beauty's memories have wide content that *entails* that it is Monday, this not an option.

I don't know whether Beauty can tell that it is Monday (on Monday), or whether her ability to tell what day it is (by memory or otherwise) is defeated. I find the defeat idea natural, but on reflection (reflection which is particularly influenced by Lasonen-Aarnio 2010, 2014), once we have granted that evidence is the sort of thing that *might* be asymmetric between Monday and Tuesday, I find it far from obvious whether it is in this case. (I suspect the answer might depend on details of Sleeping Beauty's psychological story which are omitted or differ between alternative tellings.)

If Sleeping Beauty *can* tell that it is Monday, then this straightforwardly gives us a striking answer to the question of what degree of belief in Heads is supported by her evidence on Monday. Starting from the evidence adduced in Section 1—Woke, Day, Coin, and Protocol, which supported one-third probabilities in Table 2—we add the further evidence Monday, applying the Combined Evidence principle. Thus this much of her Monday evidence supports the probabilities in Table 3, and degree of belief one half in Heads.

Table 3: Probabilities supported by Woke, Day, Coin, Protocol, and Monday

| | Monday | Tuesday |
|---|---|---|
| Heads | 1/2 | 0 |
| Tails | 1/2 | 0 |

"Halfing" is a standard position. But we arrived at this number by a different route, and the overall position that emerges is quite different from more familiar Halfer views.

First, Lewis (2001) argues for the Half answer, but he also agrees with the indifference reasoning from Elga (2000) that Monday-and-Tails and Tuesday-and-Tails

[12] The exception might be if she has further evidence with probability *zero* conditional on the fact that yesterday was Sunday. I set this complication aside.

should receive the *same* probability. So Lewis-halfers hold that Beauty's probabilities should look like Table 4, instead. But if Beauty can tell that it's Monday, then we should not accept indifference reasoning that says her evidence supports Monday-and-Tails and Tuesday-and-Tails to the same degree: for her evidence *rules out* Tuesday-and-Tails, but not Monday-and-Tails. (The indifference reasoning might be motivated by the thought that her evidence on Monday and her evidence on Tuesday are *symmetric*—but the line of reasoning we are exploring rejects this thought.) This also leads Lewis to the peculiar view that if Beauty *gains* the evidence that it is Monday, she should *raise* her degree of belief in Heads to 2/3. Our reasoning does not take us there.

Table 4: Lewis-halfing

| | Monday | Tuesday |
|---|---|---|
| Heads | 1/2 | 0 |
| Tails | 1/4 | 1/4 |

Second, consider a standard argument for Halfing (compare Lewis 2001):

(1) On Sunday Beauty's evidence supports credence one-half in Heads
(2) On Monday, Beauty's credence in Heads should be the same as her credence on Sunday.

Hawley (2013, 85) argues along these lines; he supports premise (2) with appeal to a principle he calls *Inertia*:

> If you should have degree of belief $d$ that $p$ at time $t_1$, and between $t_1$ and a later time $t_2$, your cognitive faculties remain in order, and you neither gain nor lose relevant evidence, then you should also have degree of belief $d$ that $p$ at time $t_2$.

Hawley claims, furthermore, that between Sunday and Monday, Beauty neither gains nor loses evidence relevant to the coin flip. In particular, he argues (with many others) that when Beauty learns she woke up today, this is irrelevant to the coin flip. We have rejected this. Beauty's evidence that she woke up today *is* relevant to the coin flip, given her further evidence Day, Coin, and Protocol. If one-third is not the right answer, it is because this isn't *all* of Beauty's relevant evidence (beyond that which she already had Sunday). In particular, if she also has the further evidence *It's Monday*, then this further evidence *screens off* the relevance of her waking to Heads.

Neither Woke nor Monday are independent of Heads given the rest of what Beauty knows—rather, they are exactly counterbalancing evidence.

Third, while Hawley (2013) also argues for the conclusion that Beauty should have the "half-half-zero" probabilities in table Table 3 (rather than the Lewis-halfer probabilities in Table 4), he recommends "the following *optimistic policy:* believe to degree 1 that it is Monday whenever awakened during the experiment" (p. 88). That is, Hawley recommends that Beauty should have the *same* probabilities if woken on *Tuesday*. The reasoning we have followed, though, starts from the idea that Beauty's evidence on Monday is *different* from her evidence on Tuesday (even though her predicament is phenomenally the same). On *Tuesday* she can't tell what day it is—so in this case the original Thirder reasoning remains untouched. The view we have arrived at (in contrast with Hawley's) is that Beauty's evidence depends on what day it is; on Monday her evidence supports Halfing, and on Tuesday it supports Thirding.[13]

## 3 Cosmological Evidence

An account of the Sleeping Beauty puzzle according to which her evidence is asymmetric between Monday and Tuesday, might seem a bit obtuse. You might think that the *point* of this puzzle is to explore how things go in cases where you can't tell what time it is. So the details ought to be filled in however necessary to ensure that verdict. (If informing her of the experimental protocol does not suffice to *defeat* her ordinary evidence from memory, then more direct memory-destroying measures can be taken.) Lewis, for instance, stipulates that not only are the three possible wakings in the experiment "indistinguishable," but also that "if she is awakened on Tuesday the memory erasure on Monday will make sure that her *total evidence* at the Tuesday awakening is *exactly the same* as at the Monday awakening" (2001, 171, my emphasis).[14]

Be that as it may, we also want to draw lessons from Sleeping Beauty for other cases where these details are not up for stipulation—perhaps because they are *actual* cases. Whether or not we count cases with asymmetric wide evidence as official *Sleeping Beauty* cases, if there are any cases like that, we do want to understand how things go for them.

[13] Schwarz (2012, 237) also argues for this asymmetric policy, on different grounds.

[14] This stipulation only makes sense given certain theories of the nature of temporary evidence. For example, it requires that when you can tell you're awake on Monday, this is the *same* evidence that you have when you can tell you're awake on Tuesday.

One important standard application of lessons from the Sleeping Beauty literature is in the epistemology of cosmology (see for example Bostrom 2002). According to some live cosmological hypotheses, the universe is vast—so vast that it is overwhelmingly probable that the chaotic motion of scattered atoms will coalesce into "Boltzmann brains" in the void. Moreover, this is likely to happen zillions of times, producing multitudes of Boltzmann brains with qualitative experiences just like those you are presently having. The worry is that Boltzmann brain hypotheses like these are analogous to Tails possibilities in the Sleeping Beauty—there are very many "wakings" rather than few—and so such hypotheses should receive outsized degree of belief. In that case, it seems that we should be highly confident in cosmological hypotheses according to which the universe is vast, and highly *unconfident* that we are the embodied creatures that we seem to be, rather than brains in the void.

But—whether or not Sleeping Beauty has wide evidence about what day it is—if *we* have wide evidence about our own situations, this defuses the threat of Boltzmannian skepticism. Here is a hand; here is another. If these observations are part of my evidence, then my evidence is incompatible with being a Boltzmann brain. Thus, even if my own qualitative experience *is* strong evidence that I am in a vast universe full of brains in the void,[15] the fact that I have hands tells me I am *not* such a brain in the void. This is counterbalancing evidence against such a cosmological hypothesis—leaving me right where I started on the question of the size of the universe.

(If there *are* Boltzmann brains, they have no such evidence themselves—and so their cosmological credences should accordingly be very different from mine. But Boltzmann brains also haven't really done any of the relevant science, or even read anything about it, so if evidence is wide then their credences should be *very* different from mine.)

## 4 Multifarious Evidence

So far I have treated the idea of *evidence* as a fixed point; but I am sympathetic to the view that evidence is shifty. Your evidence is what you should take for granted, which supports beliefs to various degrees. But there may be no single thing that plays this role once and for all: it is natural to think that some facts may be taken for granted for some purposes, and others for others.[16]

[15] But I am not sure whether this part of the analogy holds up. Existing isn't exactly like being awake, and I'm not sure what evidence we have which is analogous to "It's either Monday or Tuesday."

[16] For example, Greco (2017) defends this picture. Analogous views about *knowledge* rather than evidence are very widely defended (for overview see Rysiew 2016).

We began by asking what degree of belief you ought to have; but the word "ought" is notoriously context-sensitive (see for example Kratzer [1981] 2002). Maybe even when we narrow attention to what you ought to believe in an "epistemic" sense, we still haven't pinned down a single thing. If that's right, then there may be no univocal question of what degree of belief Sleeping Beauty ought to have in Heads, and likewise no univocal question of what degree of belief her evidence supports, because there is no univocal question of what her evidence *is*.[17]

If that picture is right (and I think it might be) then one plan of attack is to clearly distinguish the different candidate kinds of "evidence," and examine what attitude *each* of them supports, and then try to get clear on which notion of evidence is the relevant one for our particular purposes in a context. This goes both for the Sleeping Beauty puzzle and also for cosmological hypotheses.

## References

Arntzenius, Frank. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy* 100 (7): 356–70. http://www.jstor.org/stable/3655783.

Bostrom, Nick. 2002. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

Briggs, R. A. 2009. "Distorted Reflection." *Philosophical Review* 118 (1): 59–85. https://doi.org/10.1215/00318108-2008-029.

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago]University of Chicago Press.

Dorr, Cian. 2002. "Sleeping Beauty: In Defence of Elga." *Analysis* 62 (4): 292–96.

Elga, Adam. 2000. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60 (2): 143–47. https://doi.org/10.1093/analys/60.2.143.

———. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69 (2): 383–96. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x.

Greco, Daniel. 2017. "Cognitive Mobile Homes." *Mind* 126 (501): 93–121. https://doi.org/10.1093/mind/fzv190.

[17] I am also moved by the arguments from Srinivasan (2015) that there are no non-trivial norms (in particular, epistemic norms) which are followable in all circumstances. Then it might make sense to give *different* guiding advice about what one ought to believe, suitable for different circumstances or different purposes.

Hawley, Patrick. 2013. "Inertia, Optimism and Beauty." *Noûs* 47 (1): 85–103. https://doi.org/10.1111/j.1468-0068.2010.00817.x.

Hedden, Brian. 2015. "Time-Slice Rationality." *Mind* 124 (494): 449–91. https://doi.org/10.1093/mind/fzu181.

Keynes, John Maynard. 1921. *A Treatise on Probability*. Dover Publications.

Kratzer, Angelika. (1981) 2002. "The Notional Category of Modality." In *Formal Semantics: The Essential Readings*, edited by Hans-Jürgen Eikmeyer and Hannes Rieser, 289–323. Blackwell Oxford.

Lasonen-Aarnio, Maria. 2014. "Higher-Order Evidence and the Limits of Defeat." *Philosophy and Phenomenological Research* 88 (2): 314–45. https://doi.org/10.1111/phpr.12090.

———. 2010. "Unreasonable Knowledge." *Philosophical Perspectives* 24 (1): 1–21. https://doi.org/10.1111/j.1520-8583.2010.00183.x.

Leslie, John. 1989. *Universes*. Routledge.

Lewis, David. 2001. "Sleeping Beauty: Reply to Elga." *Analysis* 61 (3): 171–76.

Meacham, Christopher J. G. 2016. "Ur-Priors, Conditionalization, and Ur-Prior Conditionalization." *Ergo: An Open Access Journal of Philosophy* 3. https://doi.org/10.3998/ergo.12405314.0003.017.

Moss, Sarah. 2012. "Updating as Communication." *Philosophy and Phenomenological Research* 85 (2): 225–48. https://doi.org/10.1111/j.1933-1592.2011.00572.x.

Ramsey, F. P. (1926) 2010. "Truth and Probability." In *Philosophy of Probability: Contemporary Readings*, edited by Antony Eagle, 52–94. Routledge.

Ross, Jacob. 2010. "Sleeping Beauty, Countable Additivity, and Rational Dilemmas." *Philosophical Review* 119 (4): 411–47. https://doi.org/10.1215/00318108-2010-010.

Rysiew, Patrick. 2016. "Epistemic Contextualism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/contextualism-epistemology/.

Salow, Bernhard. 2018. "The ExternalistGuide to Fishing for Compliments." *Mind* 127 (507): 691–728. https://doi.org/10.1093/mind/fzw029.

Schwarz, Wolfgang. 2012. "Changing Minds in a Changing World." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 159 (2): 219–39.

http://www.jstor.org/stable/23262286.

Siegel, Susanna. 2012. "Cognitive Penetrability and Perceptual Justification." *Noûs* 46 (2).

Srinivasan, Amia. 2015. "Normativity Without Cartesian Privilege." *Philosophical Issues* 25 (1): 273–99. https://doi.org/10.1111/phis.12059.

Stalnaker, Robert C. 2008. *Our Knowledge of the Internal World*. Oxford University Press.

Sullivan, Meghan. 2014. "Change We Can Believe in (and Assert)." *Noûs* 48 (3): 474–95. https://doi.org/10.1111/j.1468-0068.2012.00874.x.

Titelbaum, Michael G. 2013. "Ten Reasons to Care About the Sleeping Beauty Problem." *Philosophy Compass* 8 (11): 1003–17. https://doi.org/10.1111/phc3.12080.

van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford University Press.

Weatherson, Brian. 2011. "Stalnaker on Sleeping Beauty." *Philosophical Studies* 155 (3): 445–56.

Weintraub, Ruth. 2004. "Sleeping Beauty: A Simple Solution." *Analysis* 64 (281): 8–10. https://doi.org/10.1111/j.0003-2638.2004.00453.x.

Williamson, Timothy. 2002. *Knowledge and Its Limits*. Oxford University Press. https://doi.org/10.1093/019925656X.001.0001.