# Sleeping Beauty: Why Everyone Should Be a Thirder

Lennart Ackermans

11th February 2024

**Abstract**

The last two decades have seen a heated debate between "halfers" and "thirders": those who believe Sleeping Beauty's credence in a coin landing heads is $1/2$ and those who believe it is $1/3$ – as well as quite some alternative positions. This paper attempts to put an end to halfism. I present a new argument for thirdism which cannot be resisted using any of the previously used halfer strategies. My argument uses an analogy in which Sleeping Beauty has a lucid dream on each day. To arrive at thirdism, she uses an unproblematic type of Bayesian conditionalisation, the principle of indifference, and the principal principle. I argue that all of these reasoning steps should be uncontroversial. Finally, I argue that all published defences of halfism are untenable.

## 1 Introduction

Sleeping Beauty, a genius of rationality, participates in an experiment at the Experimental Philosophy Lab. On Sunday evening, she is told how the experiment will go and is put to sleep. The researchers then toss a fair coin. On Monday afternoon, Beauty will be briefly woken up, interviewed by the researchers, and put to sleep again. Beauty will be woken up again on Tuesday afternoon if and only if the coin came up tails. However, the researchers will also give Beauty an amnesia-inducing drug on Monday evening that makes her forget everything that happened on Monday. Hence, if she wakes up again on Tuesday, her available evidence is identical to her Monday awakening. On Wednesday, Beauty is woken up for the last time, and the experiments ends. One further detail is often added: a

little while into Beauty's Monday interview (but not her Tuesday interview), the researchers tell her it is Monday.

Beauty knows exactly how the experiment will go. What is Beauty's credence that the coin lands heads when she is interviewed on Monday afternoon, but before she is told that it is Monday? If there is a unique rational credence, then Beauty (a rationality genius) will choose it. Surprisingly, there is much disagreement about the answer to this question. There are two main camps in this debate. The "thirders" maintain that Beauty's credence is 1/3. The "halfers" claim that it is 1/2.

Since the question was first posed in print by Elga (2000) – a thirder – and responded to by Lewis (2001) – a halfer – a vigorous debate ensued that is still raging on today. The arguments on both sides have become increasingly complex in response to attacks from the other side. While thirdism seems to be the majority opinion, there is no consensus in sight about the solution to the problem nor about the source of the disagreement.

Like many previous authors, I believe thirdism is simply correct and halfism wrong. I also believe that early arguments by Elga (2000), Horgan (2004) and others already established this. Apparently, however, these arguments have left open too much space for halfers (and others) to dissent. I attempt to set this right by offering yet another argument for thirdism. This argument, I maintain, cannot be resisted using any of the existing strategies used by halfers.

My argument is similar to Horgan's, but instead of synchronic Bayesian updating, it uses a less controversial diachronic principle of Bayesian updating. My argument works by constructing an analogous case in which Beauty has a lucid dream on both Monday and Tuesday before she is woken up. After being woken up on Monday, she updates her credences based on her dreaming credences. My argument also bears resemblance to other arguments by analogy, such as Dorr (2002) and Arntzenius (2003), but my lucid dreaming case is more clearly analogous to the original problem than previously offered analogies.

The primary benefit of my argument is that it makes very clear why halfer objections to it do not succeed. A halfer would either have to object to Beauty's credences while dreaming, or object to her use of Bayesian updating. The first strategy does not succeed both because these credences are highly plausible and because the type of reasoning typically used to defend halfism supports these credences. The second strategy does not succeed because Beauty's use of updating, I will argue, ought to be uncontroversial (even though it involves so-called *self-locating beliefs*). Since the use of such updating is *prima facie* rational, halfers would need to point to some problem with its use in this particular case; but I argue

that problems identified in the literature do not apply to this case, and that further problems are unlikely to exist.

After presenting my positive case for thirdism, I turn to its two primary contenders, called *Lewisian halfism* and *double halfism*. Using my findings from the earlier sections, I argue that both positions are untenable. By the end of the article, I have refuted arguments from a number of halfers, including Lewis (2001), Bostrom (2007), Pust (2008, 2012), Meacham (2008) and Bradley (2011, 2012).

This article cannot address all arguments against thirdism. In particular, I avoid addressing positions different from both halfism and thirdism, such as the position that Beauty should have imprecise credences (Singer 2014) and positions that her credence should have a value in between 1/2 and 1/3 (Cisewski et al. 2016). I also avoid addressing analyses of the problem which do not model Beauty's self-locating beliefs (or *centred beliefs* or *indexical beliefs*).[1]

The structure of the article will be as follows. Section 2 introduces the analogous case, introduces an uncontroversial principle of Bayesian updating, argues for thirdism in the analogy, and defends the analogy. In section 3, I discuss various objections in the literature to thirdism and argue that they fail as objections to my argument. This concludes my positive argument for thirdism. In section 4, I argue that halfism is untenable. Section 5 concludes and gives some reasons to be sceptical of the intuitions leading to halfism.

## 2 Beauty's Lucid Dreams

I will use the following analogy to the Sleeping Beauty problem. The experiment is performed in the same way as the original, but instead of its usual test subject Sleeping Beauty, the experiment is performed on Lucid Beauty. Lucid Beauty, like Sleeping Beauty, is always entirely rational. Additionally, Lucid Beauty has a dream every morning when she is sleeping in the Experimental Philosophy Lab. These dreams are lucid dreams: she knows that she is dreaming. Moreover, Lucid Beauty

1. An example of the latter is Cisewski et al. (2016). In the authors' modelling, Beauty's new evidence after being awoken on Monday is that she has the particular experience $x$ of being woken up on either Monday or Tuesday. This evidence is non-self-locating, since it is the disjunction of the proposition that she experiences $x$ on Monday and the proposition that she experiences $x$ on Tuesday. Beauty uses Bayesian updating on this evidence based on her Sunday probability function. Notably, the authors do *not* model the self-locating evidence that Beauty is woken up *today*. While I do not address the issue in this article, I believe that models which do not express the self-locating aspect of Beauty's evidential situation are insufficiently strong.

| | Sunday | Monday | Tuesday | |
|---|---|---|---|---|
| | | | Heads | Tails |
| Morning | | Lucid dream | Lucid dream | |
| Afternoon | | Beauty woken up | Asleep | Woken up |
| Evening | Experiment explained<br>Put to sleep | Put to sleep<br>Drug administered | Put to sleep | |

**Figure 1:** Overview of the Lucid Beauty experiment.

remains entirely rational during her lucid dreams, and she has access to all the information about the experiment that she is given on Sunday evening.

Like in the original experiment, Beauty is given an amnesia-inducing drug on Monday evening. Hence, her experience during her Monday dream is identical to her Tuesday dream. Like in the original problem, Beauty is only woken up on Tuesday afternoon if the coin landed tails. See figure 1 for a schematic overview.

Beauty's credences are modelled by a credence function $P$. The domain of this credence function is a set of propositions for which Beauty has credal attitudes. The set of propositions which Beauty believes to be true are called her *evidence*.

Lucid Beauty evidence's during her Monday morning dream is given by the evidence {LUCID, MON ∨ TUE, FAIR, CONDITIONS} described below. When she is woken up in the afternoon, AWAKEN is added to her evidence. Her evidence then consists of the following.

LUCID: I have a lucid dream on both Monday morning and Tuesday morning.

MON ∨ TUE: Today it is either Monday (MON) or Tuesday (TUE).

FAIR: A fair coin was tossed.

CONDITIONS: The researchers awaken me today if and only if either (MON) it is Monday or (TAILS) the coin landed tails.[2]

AWAKEN: The researchers awaken me today.[3]

What is Lucid Beauty's credence that the coin lands heads after she is woken

2. This sentence correctly describes the conditions only given Beauty's evidence MON ∨ TUE. Without that evidence, CONDITIONS should state: "If today is either Monday or Tuesday, then the researchers awaken me today if and only if it is Monday or the coin landed tails."

3. The evidence AWAKEN is tenseless to emphasize that "The researchers will awaken me today" and "The researchers have awakened me today" are effectively the same evidence, that is, they can both be treated as equivalent to AWAKEN by Beauty. Hence, Beauty can entertain the proposition AWAKEN both during her lucid dream (when she is still unsure of its truth) and after she is awakened.

up by the researchers on Monday afternoon? I argue that it is 1/3. I then argue that Lucid Beauty is analogous to Original Beauty, such that Original Beauty also assigns 1/3 to the coin landing heads. First, I need to introduce a central premise of my argument, which is that Lucid Beauty can use a version of Bayesian updating to conditionalise on AWAKEN.

## 2.1 Bayesian Updating for Self-Locating Evidence

It is often argued that traditional Bayesian updating is unavailable for self-locating beliefs (e.g., Meacham 2008; Pust 2012). Self-locating beliefs are beliefs about who you are or when you are. For example, the belief that today is 1 January 2023 locates one's current self at a particular moment in time. Many sentences which contain indexicals such as "I" and "Today" signify self-locating beliefs (while not all of them do). For example, Beauty's belief, on Monday, that she was just woken up by the experimenters is self-locating. In this scenario, Beauty's "just" does not indicate an exact moment in time, since it can either be Monday or Tuesday. (*We* know that "just" is on Monday, but Beauty does not. Thus, the belief that she was woken up on Monday is different from the belief that she was "just" woken up, and only the latter belief is self-locating.) The belief can be said to locate Beauty's self (as it exists at when she utters "just") at the particular moment in time when she has in fact been woken up by the experimenters.

It is easy to see how conditionalisation on self-locating evidence *can* be problematic. Suppose that at 12:00 I know that it is 12:00, and my credence function at 12:00 is $P_{12:00}$. At 12:05 I learn that it is 12:05. Traditional Bayesian updating would suggest creating a new credence function $P_{12:05}(\cdot) = P_{12:00}(\cdot \mid \text{It is 12:05})$. However, the statement "It is 12:05" was false at 12:00, and conditioning on a statement with a credence of 0 gives undefined credences. Hence, Bayesian updating seems to be unavailable in this situation.

However, this does not imply that conditioning on self-locating evidence is always problematic. If at 12:00 I'm not sure which day it is, and at 12:05 I learn that it is Monday, no clear counterexample for Bayesian updating such as the above exists. Bradley (2011) argues that there are cases in which traditional Bayesian updating on self-locating evidence is unproblematic and delivers plausible results. Moreover, several proposals have been made that extend Bayesian updating to self-locating beliefs (e.g., Titelbaum 2008; Schwarz 2012; Schulz 2010).

Proposals to save Bayesian updating for self-locating evidence typically do two things. First, they determine a subset of learning situations in which the tradi-

tional principle of Bayesian conditionalisation can be applied to self-locating beliefs without making any fundamental changes to the principle. Roughly speaking, these are situations in which (the agent knows that) the self-locating proposition's truth value is unchanged between $t_1$, the time of the agent's prior credence function, and $t_2$, the time at which the self-locating belief is adopted. Call this *type 1 updating*. Second, these proposals attempt to extend Bayesian conditionalisation to learning situations which do not clearly fall into the first category. Call this *type 2 updating*.

Since my argument's use of conditionalisation falls into the first, less controversial, category, I will not be concerned with type 2 updating. Nevertheless, objections have been made even against type 1 updating. One type of objection claims that counterintuitive results can be obtained by using Bayesian conditionalisation on self-locating evidence; I will turn to this type of objection in section 3.3. A second objection is that, on at least some accounts of what beliefs and propositions are, traditional Bayesian conditionalisation is impossible (Pust 2012). In this section, I will give an account of type 1 Bayesian updating on self-locating evidence that I argue defies Pust's objections. The main purpose of this discussion is to argue that type 1 Bayesian updating as I describe it is just *normal* Bayesian updating – and thus, should be given the same plausibility as normal Bayesian updating. I will return to Pust's objection that such a thing is impossible in section 3.2.

Pust claims that any kind of conditionalisation on a newly acquired self-locating belief on the basis of an earlier credence function is impossible, because the new belief could not have been in the domain of the credence function at the earlier time. He argues for this claim three times – each using a different account of what an indexical belief is (but the arguments are very similar in structure). Here I adopt Pust's first account, which retains the traditional view that beliefs are propositions entertained by a person. The benefit of this approach is that it is the most simple and well-known. It allows us to keep thinking of the elements of the domain of a credence function as propositions. Self-locating beliefs such as "today is Monday" can simply be thought of as propositions with a temporal indexical called *temporally indexical propositions*.

This account attempts to keep most of the tenets that are traditionally associated with propositions; in particular, that each proposition has only one truth value, independently of time and place. It follows from this that a sentence such as "It is raining now", uttered at 12:00, cannot signify the same proposition as that sentence uttered at 12:05. Hence, to identify a unique proposition that is temporally indexical, one needs both the sentence and the time at which it is entertained. For example, the two propositions just discussed are:

$\textsc{rain}_{12:00}$: It is raining now.
$\textsc{rain}_{12:05}$: It is raining now.

I use the notation $\textsc{x}_t$ to refer to the proposition described by the sentence X, entertained at $t$. (Later, the notation X will refer to a proposition. In this section, non-indexed small caps words are sentences and indexed small caps words are propositions.)

As an aside, indexical sentences might sometimes refer to non-indexical propositions, and it's important to understand the difference. For example, take the sentence "Yesterday was the rainiest day in the month." By "Yesterday", the speaker of this sentence might have meant "May 20, 2023," in which case the proposition is non-indexical. Hence, this sentence has two interpretations, an indexical and a non-indexical proposition.

Pust claims that the account sketched above requires one to accept that there are propositions of *limited accessibility*. This claim relies on the distinction between *entertaining* (or *employing* or *expressing*) and *referring to* (or *thinking about*) a proposition. Contemplating the truth of $\textsc{rain}_{12:00}$ is to entertain the proposition – and one typically does so only at 12:00. On the other hand, describing the proposition itself only requires us to refer to it, as I do with the notation 'X'. If $\textsc{rain}_{12:00}$ is a proposition of limited accessibility, it can *only* be entertained at 12:00. At 12:05, the same proposition can no longer be entertained, although one can entertain another proposition $\textsc{rain}_{12:05}$.

I believe there may indeed be propositions of limited accessibility. However, Pust make the more controversial claims that all temporally indexical propositions are propositions of limited accessibility, and that they are accessible *only at a single point in time*. Hence, Pust would have to deny that the sentence "It is raining now" entertained at 12:05 accesses the same proposition as the sentence "It will rain in 5 minutes," entertained at 12:00. If this is so, it is seemingly impossible for an agent who receives the evidence $\textsc{rain}_{12:05}$ to conditionalise using her earlier credence function – at which time this proposition was inaccessible, and thus could not have had a credence.

The position that all temporally indexical propositions are accessible at only a single point in time I will call *extreme indexicalism*. In section 3.2, I argue that even when accepting extreme indexicalism, there is a way to save Bayesian updating. Here, I will propose an alternative to extreme indexicalism that I find more natural.

Extreme indexicalists would have to hold that the referents of temporal indexicals are always *points* in time. Hence, even a sentence containing the indexical "Today" would signify a different proposition when it is entertained at different

moments on Monday. For example, returning to the setting of Sleeping Beauty, according to extreme indexicalism, the following two propositions are different propositions:

$\text{MON}_{\text{Mon 12:00}}$: Today is Monday.
$\text{MON}_{\text{Mon 12:05}}$: Today is Monday.

(I don't know whether Pust subscribes to this exact version of extreme indexicalism. What matters it that my arguments below succeed if it can be rejected.)

There is an alternative view which seems more plausible. Instead of a point in time, an indexical could refer to a time range. For example, the indexical *today* could refer to the range of time between 00:00 and 24:00 on the day of entertaining the proposition. The proposition $\text{MON}_{\text{Mon 12:00}}$ would state that this time range is a Monday. Clearly, the time range that *today* refers to is the same at 12:00 and 12:05. Hence, $\text{MON}_{\text{Mon 12:00}}$ and $\text{MON}_{\text{Mon 12:05}}$ are in fact the same proposition, and both propositions are more naturally described as $\text{MON}_{\text{Mon}}$ (the proposition accessed on Monday by the sentence MON). Since the sentence MON signifies the same proposition, entertained at different times on Monday, it follows that this proposition is accessible at multiple moments in time.

It can now be seen that type 1 Bayesian updating is possible and is just traditional Bayesian updating. Traditional Bayesian updating between $t_1$ and $t_2$ can straightforwardly be applied if two conditions are satisfied. First, all certainties at $t_1$ remain certainties at $t_2$. That is, for each proposition X in the agent's credence domain at $t_1$, we have $P_1(\text{X}) = 1$ only if $P_2(\text{X}) = 1$. Second, the credence domain at $t_1$ equals the credence domain at $t_2$. When there are temporally indexical propositions in the credence domain, this means the following. Each proposition $\text{X}_{t_1}$ entertained by the agent at $t_1$ is the same proposition as the proposition $\text{X}_{t_2}$ entertained by the agent at $t_2$, where X is a possibly indexical sentence; and each proposition $\text{X}_{t_2}$ entertained by the agent at $t_2$ is the same as the proposition $\text{X}_{t_1}$ entertained by the agent at $t_1$.

For example, if an agent's credence domain at 12:00 consists only of $\text{MON}_{12:00}$ and some non-indexical propositions, and her credence domain at 12:05 consists only of $\text{MON}_{12:05}$ and the same non-indexical propositions, the second condition is satisfied. After all, $\text{MON}_{12:00}$ equals $\text{MON}_{12:05}$. Suppose she is unsure whether it is Monday at 12:00 and learns that it is Monday at 12:05, while her other evidence does not change between 12:00 and 12:05. Both conditions are now satisfied, and as is generally agreed, this allows the agent to use traditional Bayesian updating. That is, she can set $P_{12:05}(\cdot) = P_{12:00}(\cdot \mid \text{MON}_{12:00})$.

An important caveat is that in order to know that she can use Bayesian updating, the agent must know that the second condition is satisfied. This is not always the case: for example, suppose that at 23:00 the agent knows that is Monday 23:00. After some time has passed without looking at the clock, she should become uncertain whether the proposition "Today is Monday," entertained at that time, is the same as the proposition "Today is Monday," entertained at 23:00. In this case, traditional Bayesian updating seems unavailable. (But Schulz 2010 extends Bayesian updating to situations similar to this one.)

Putting this together, we can stipulate conditions under which an agent should uncontroversially use traditional Bayesian conditionalisation on self-locating evidence as follows.

> **Type 1 Bayesian updating for self-locating evidence.** Let $\Omega_t$ be an agent's credence domain (a set of propositions) at $t$. Suppose that the agent's credence domain at $t_1$ and $t_2$ is generated by a set of sentences $S$, that is, we have $\Omega_t = \{\mathrm{S}_t \mid \mathrm{S} \in S\}$, for $t \in \{t_1, t_2\}$. Finally, suppose that the credence domains are identical and that the agent knows this, i.e. the agent knows that $\Omega_{t_1} = \Omega_{t_2}$. Then we have $P_2(\mathrm{X}_{t_2}) = P_1(\mathrm{X}_{t_2} \mid \mathrm{Y}_{t_1})$, for all $\mathrm{X} \in E$.

In the following sections, when the sentences generating Beauty's credence domain at different times signify the same propositions, I will leave out the subscripts.

## 2.2 A Derivation of Thirdism for Lucid Beauty

Let $P$ be Beauty's credence function after the researchers awaken her in the afternoon. Let $P_*$ be her credence function during her lucid dream on Monday morning. In later sections, $P_-$ will denote Beauty's credence function on Sunday evening, and $P_+$ will denote her credence function after the researchers tell her it is Monday.

The following premises are all that is needed for my derivation of thirdism. I will defend these premises in the next subsection.

P1. $P_*(\textsc{mon}) = 1/2$.

P2. $P_*(\textsc{heads}) = 1/2$.

P3. HEADS and MON are independent given Beauty's evidence during her dream (LUCID, MON $\vee$ TUE, FAIR, and CONDITIONS). That is, $P_*(\textsc{heads} \,\&\, \textsc{mon}) = 1/4$.

P4. Beauty always assigns the same credence to propositions that are logically equivalent given her evidence, and she assigns the same conditional credence to propositions that are logically equivalent given her evidence and the condition.

P5. Beauty's credence functions satisfy the probability axioms.[4]

P6. Beauty determines $P$ from $P_*$ using type 1 Bayesian updating for self-locating evidence. That is, $P(\text{HEADS}) = P_*(\text{HEADS} \mid \text{AWAKEN})$.

For my derivation of thirdism, it is essential that the proposition AWAKEN is logically equivalent to MON $\vee$ TAILS given Beauty's evidence while dreaming. In appendix A.1, I show that it follows straightforwardly from this fact and the premises P1–P5 that $P_*(\text{HEADS} \mid \text{AWAKEN}) = 1/3$. After applying type 1 Bayesian updating (P6) we get $P(\text{HEADS}) = 1/3$.

## 2.3 Defending the Premises

I will assume that P4 and P5 will not be challenged, given that they have not been challenged in the Sleeping Beauty literature before and that there seems no good reason to challenge them. That leaves P1, P2, P3 (the prior probabilities) and P6 (type 1 Bayesian updating) to give a positive defence for.

I will first give a preliminary defence of P1, P2 and P3 together. Consider what Beauty's credences during her dream would be if she did not know CONDITIONS, that is, if she knew only {LUCID, MON $\vee$ TUE, FAIR}. Call her probability function in this situation $P_*^0$. Given this evidence, I would expect most readers to agree that she assigns a credence of $1/2$ to both MON and HEADS, and that she would consider the two independent. (But I will defend this below.) I claim that relative to this evidential situation, CONDITIONS is irrelevant evidence for HEADS and MON. Hence, we have $P_*^0(\text{HEADS}) = P_*(\text{HEADS}) = 1/2$, $P_*^0(\text{MON}) = P_*(\text{MON}) = 1/2$ and $P_*^0(\text{HEADS} \,\&\, \text{MON}) = P_*(\text{HEADS} \,\&\, \text{MON}) = 1/4$.

It seems intuitively obvious that CONDITIONS is irrelevant, as can be seen from the following similar case. Suppose you are unsure whether it is Monday. I toss a coin without showing you the result and then tell you that I will raise my hand a minute from now if and only if the coin landed tails or it is Monday. Before this minute has passed, this knowledge of the conditions under which I will raise my

4. $P(\text{T}) = 1$ for every tautology T; $P(\text{X}) \geq 0$ for all X; $P(\text{A} \vee \text{B}) = P(\text{A}) + P(\text{B})$ for all A and B that are mutually exclusive.

hand should clearly not affect your credences in the way the coin lands, nor in which day it is. This is the case as long as you have no additional evidence about whether I will raise my hands. (As soon as I do raise my hand, this situation clearly changes.)

This intuitively plausible reasoning seems to depend on a principle like the following.

> **Logical Irrelevance.** Suppose a rational agent is ignorant with respect to a proposition X, that is, she has no information (evidential or non-evidential) about the proposition except for (C) "Necessarily, X if and only if Y." Then the evidence C is irrelevant with respect to Y.

This principle seems plausible. Knowing that two propositions are equivalent cannot change one's credences in either proposition by itself. Only if the agent knows something about one proposition, this knowledge can be used to bear on the second via the logical equivalence.

Returning to Lucid Beauty, it seems that the only evidence that Beauty has about AWAKEN is given by CONDITIONS. Hence, except for CONDITIONS, she is ignorant of AWAKEN. By Logical Irrelevance, CONDITIONS is irrelevant for MON∨TAILS. Clearly, it must then also be irrelevant for MON and for TAILS. Hence, Beauty's credences in HEADS and MON should be the same whether or not she knows CONDITIONS. Premises P1–P3 follow.

It is left to show that the credences $P_*^0$ have the right values. (Recall, $P_*^0$ gives Beauty's credences supposing she does not know CONDITIONS.) First, $P_*^0(\text{MON}) = 1/2$ can be defended using a classical principle of indifference. According to a classical principle of indifference, if two propositions are indistinguishable (or "symmetrical") with respect to the evidence, they should be assigned the same credence. Without knowing CONDITIONS, MON and TUE are quite clearly indistinguishable. Beauty knows she will have a lucid dream on both Monday and Tuesday and that she now has one of these lucid dreams. With respect to this evidence, MON and TUE perform a symmetrical role, that is, they are indistinguishable with respect to the evidence. Moreover, they form a partition with respective to her evidence (it is either Monday or Tuesday), so their credences sum to 1. Hence, according to the classical principle of indifference, Beauty should assign a credence of $1/2$ to both MON and TUE.

The principle of indifference is controversial, since it sometimes gives conflicting recommendations for the same problem. However, there is no reason to assume that the principle is problematic in our case. Consider the following slightly different example in which its application is plausible. Take a rational agent who

lives in a world in which the day of the week is of very little relevance. Since this agent never checks which day it is, he is completely ignorant of that fact. If this person were told that it is either Monday or Tuesday, the principle of indifference recommends, plausibly, that his credence in it being Monday is $1/2$. Someone who agrees that the principle of indifference is plausibly applied in this situation, should also agree that it can plausibly be applied by Beauty during her lucid dream.

Note that P1 can also be defended directly using Elga's restricted principle of indifference (Elga 2004). This principle states that within a possible world, each location in that world at which the agent may be should receive the same credence, supposing that her subjective experience would be the same at each location. This principle is usually formulated to be about *centred worlds*. A *centred world* is a 3-tuple of an agent, a moment in time and a traditional possible world. For example, if $H$ is a traditional possible world in which the coin lands heads, Beauty would assign credences to the centred worlds $H_m = (\text{Beauty, Monday morning}, H)$ and $H_t = (\text{Beauty, Tuesday morning}, H)$. Two centred worlds are *subjectively indistinguishable* if the agent's experience in one is identical to the other.

> **Elga's restricted principle of indifference.** Centred worlds associated with the same traditional possible world, that are subjectively indistinguishable, should be assigned equal credences.

During Beauty's lucid dream, $H_m$ and $H_t$ are subjectively indistinguishable. A credence in a proposition such as MON & HEADS is equal to the credence in all centred worlds in which it is Monday and the coin lands heads. Hence, by Elga's restricted principle of indifference and the law of total probability, we have

$$P_*(\text{MON} \,\&\, \text{HEADS}) = \sum_H P_*(H_m) = \sum_H P_*(H_t) = P_*(\text{TUE} \,\&\, \text{HEADS}),$$

where the summations range over all worlds $H$ in which the coin lands heads. By a similar argument, we have $P_*(\text{MON} \,\&\, \text{TAILS}) = P_*(\text{TUE} \,\&\, \text{TAILS})$. Applying the law of total probability, we get $P_*(\text{MON}) = P_*(\text{TUE}) = 1/2$.

Moving on to P2, for which we need to show $P^0_*(\text{HEADS}) = 1/2$. This follows from a special case of Bayesian updating: only relevant new evidence may change one's probabilities. (This principle is sometimes invoked by halfers, incorrectly, to argue that $P(\text{HEADS}) = 1/2$ because $P_{\_}(\text{HEADS}) = 1/2$.[5]) With respect to Sunday, Beauty does not seem to have gained new relevant evidence: she has lost the Sunday

5. The halfers' use is incorrect, because AWAKEN, which given Beauty's evidence is equivalent to MON ∨ TAILS, is relevant to HEADS. See also section 5.1.

evening belief "Today is Sunday" and gained a new belief "Today is either Monday or Tuesday," which is clearly not a relevant evidential change (in the absence of the additional information CONDITIONS and AWAKEN). Hence, Beauty should leave her credence in HEADS unchanged from Sunday evening. As everyone in the debate agrees, Beauty's credence in HEADS on Sunday evening is $1/2$. Hence, $P^0_*(\text{HEADS}) = 1/2$ follows from an uncontroversial type of Bayesian updating.

Premise P2 can also be defended by direct application of the principal principle. According to the principal principle, one's credence in HEADS should be the objective chance of heads if one knows the chance and one has no inadmissible evidence. Here *Inadmissible evidence*, roughly speaking, is evidence that would require you to change your credence in HEADS if you learned it after having calibrated to the chance. On Sunday, everyone agrees, Beauty has no inadmissible evidence with respect to HEADS. On Monday morning, Beauty loses the evidence SUN and gains the evidence MON $\vee$ TUE and CONDITIONS. SUN and MON $\vee$ TUE are clearly irrelevant for HEADS, and CONDITIONS is irrelevant by Logical Irrelevance. Irrelevant evidence is certainly inadmissible. Hence, we can apply the principal principle, which yields P2.

For P3, we need to show that HEADS and MON are independent with respect to $P^0_*$. By themselves, these propositions are clearly unrelated, so there is dependence if and only if Beauty's other evidence connects the two. Except for CONDITIONS, Beauty has no evidence that relates HEADS and MON. Hence, they must be independent.

Lastly, turn to P6. First, consider that the conditions of type 1 Bayesian updating are satisfied: between Beauty's lucid dream on Monday morning and her awakening on Monday afternoon, her credence domain remains the same. At both times, there are sentences generating propositions in her credence domain that contain the indexical *today*. Since no day has passed, these sentences signify the same proposition. Beauty knows that no day has passed (after all, she has a lucid dream every morning), so she knows that these sentences signify the same proposition in the morning and in the afternoon. Hence, she can use type 1 Bayesian updating for self-locating evidence.

It might be objected that an additional argument is needed that Bayesian conditionalisation is appropriate in this particular situation (which involves self-locating beliefs). However, traditional Bayesian updating is a well-regarded principle of rationality with a longstanding research tradition. Hence, Bayesian conditionalisation is the *prima facie* rational way to update after obtaining new beliefs. Moreover, there is no *prima facie* reason to assume that type 1 Bayesian updating on self-

locating evidence is inappropriate, since – as I've argued in section 2.1 – type 1 updating is just the normal way of applying the traditional principle of Bayesian updating. Hence, P6 deserves *prima facie* acceptance. That said, there are specific objections to P6 that might be thought to override our *prima facie* acceptance. I turn to these objections in section 3.

## 2.4 Defending the Analogy

I argue that the version with Lucid Beauty and the original Sleeping Beauty are analogous in the following sense.

> ANA. If Lucid Beauty assigns 1/3 to HEADS, then Sleeping Beauty assigns 1/3 to HEADS in the original experiment.

If, on two separate occasions, an agent's relevant evidence is identical, and there is only one rational credence in the first case, then there is only one rational credence in the second. Let's call this the evidence transfer principle, which is a defining feature of probabilistic rationality. Principles of rationality restrict which credence you can have *given* your evidence. Hence, if the evidence permits only one credence, all rational credence functions based on identical evidence must concur. By the evidence transfer principle, someone who denies ANA has to claim that Lucid Beauty's evidence is relevantly different from Sleeping Beauty's.

It seems that the only additional evidence that Lucid Beauty has with respect to Original Beauty, after being awakened by the researchers, is LUCID, that she has a lucid dream on Monday and Tuesday morning. (Note that Lucid Beauty's remembrance of having a dream in the morning does not constitute more additional evidence, since it is logically implied by LUCID.) Hence, if indeed LUCID is the only additional evidence, then denying ANA requires claiming that LUCID is relevant evidence for heads given Beauty's other evidence after being awakened. However, it is clear that having a dream does not ordinarily provide evidence for the result of coin tosses. Moreover, LUCID does not seem to be relevant for HEADS via Beauty's other evidence: LUCID provides no evidence for the current day, since Beauty has a lucid dream on both Monday and Tuesday. Besides its reference to Monday and Tuesday, there are no other ways in which LUCID could connect to Beauty's other evidence. Hence, LUCID is not relevant evidence.

Note also that many descriptions of the original experiment do not mention that Beauty has no dreams – in particular, both Elga (2000) and Lewis (2001) make no mention of dreams or an absence thereof. It is possible, then, that Original Beauty has lucid dreams. Hence, someone who denies both thirdism in the original

experiment and ANA (but accepts my derivation of thirdism for Lucid Beauty) has to claim that Original Beauty's credence in HEADS is indeterminate, where the exact credence depends on additional facts not mentioned in the original description of the experiment. This would constitute a major departure from halfism.

A final way to deny ANA is to claim that I have left out evidence, besides LUCID, that is relevant, that Lucid Beauty has, but that Original Beauty doesn't have. Given the extreme similarity between the two cases, this approach seems unlikely to succeed.

# 3 Objections

## 3.1 'AWAKEN Is Not New Information'

It is sometimes claimed that AWAKEN is not new information. If this is so, Beauty's Monday afternoon credence in HEADS should remain the same as during her lucid dream, as well as her credence on Sunday evening, which both are $1/2$. I show that two existing arguments that AWAKEN is not new information fail, at least when applied as objection to my argument.

One way in which it has been argued that Beauty gains no new information when she is woken up is to say that Beauty already knew on Sunday evening that she would awaken at least once. If she then wakes up on Monday, she learns nothing knew: she was already certain she would wake up. (See Lewis 2001 and Bradley 2012 for versions of this argument.) It can be seen that this argument fails in two ways. First, it works by understanding the belief of being awakened as the non-indexical proposition "There is a day at which I am awakened before the end of the experiment," rather than the temporally indexical proposition AWAKEN: "The experimenters awaken me today." Beauty is indeed certain at all times of the former proposition; but during her lucid dream, she is uncertain of the latter. Hence, after she is awakened, compared to her dream, Beauty has new relevant evidence. Second, when Beauty is awakened on Monday, this awakening could be her second, for all she knows. Hence, by being awakened, she not only learns that she is awakened at least once (which she already knew), but she also receives information about the possibility of a second awakening. Hence, her new evidence also ought to change her non-indexical beliefs.

A more sophisticated argument that AWAKEN is not new information is offered by Pust (2008), who attempts to undermine the argument for thirdism by Horgan (2004, 2007), which is similar to mine. Horgan's argument involves *synchronic*

*Bayesian updating*, a type of Bayesian updating done at a single moment in time instead of at a later time based on an earlier credence function. In Horgan's reasoning, after Beauty is woken up by the experimenters, she imagines what her credence function would be if she had the belief she has without AWAKEN, that is, if her evidence were given by {MON ∨ TUE, FAIR, CONDITIONS}. This gives her a preliminary credence function $P_1$. She then creates her actual credence function $P_2$ using Bayesian updating on AWAKEN. Her credence in HEADS becomes $P_2(\text{HEADS}) = P_1(\text{HEADS} \mid \text{AWAKEN})$.

Pust argues that Beauty's preliminary credence in HEADS & TUE is 0, while Horgan argues it is 1/4. Horgan's preliminary credences are the same as Lucid Beauty's priors while dreaming in my argument, given by P1–P3. Similarly to my argument, Horgan's arguments concludes with $P_2(\text{HEADS}) = 1/3$.

Pust's counterargument is as follows, where H2 equals the proposition HEADS & TUE.

> (1) An epistemic probability is the degree to which an agent in some logically possible epistemic situation ought (rationally) to believe some statement.
>
> (2) Any logically possible agent in any logically possible epistemic situation ought to be absolutely certain that the statement "I am conscious now" is true.
>
> (3) Thus, (when she is awake on Monday) Beauty's preliminary probability for "I am conscious now" is one. [1, 2]
>
> (4) Beauty's preliminary probability for "I am conscious now only if I am awakened today by the experimenters" is one.
>
> (5) Thus, Beauty's preliminary probability for "I am awakened today by the experimenters" is one. [3, 4]
>
> (6) Beauty's preliminary probability for H2 with respect to the statement "I am awakened today by the experimenters" is zero.
>
> (7) So, Beauty's preliminary probability for H2 is zero. [5, 6] (Pust 2008, 99)

(Horgan 2008 gives his own rebuttal, to which Pust 2013 replies. My rebuttal is different.)

Note that Pust's argument works by deriving first that "I am awakened today by the experimenters" (AWAKEN) is not new information (premise 5). This would seem to undermine both my argument and Horgan's. In my argument, this would

mean that Beauty's credence after being woken up by the experimenters, $P(\text{HEADS})$, is the same as her credence during her lucid dream, $P_*(\text{HEADS}) = 1/2$.

However, Pust's counterargument clearly does not work if it is changed to apply to my argument. Instead of synchronic Bayesian updating on a preliminary credence function, Lucid Beauty uses standard Bayesian updating on her dreaming credences $P_*$. Hence, premises 4 and 5 become:

(4) Beauty's dreaming probability for "I am conscious now only if I am awakened today by the experimenters" is one.

(5) Thus, Beauty's dreaming probability for "I am awakened today by the experimenters" is one. [3, 4]

It is obvious that these premises are false. Beauty is conscious during her lucid dream. Hence, it is clear that her credence in the statement from premise 4 is $P_*$(I am conscious now only if I am awakened today by the experimenters) = 0. It is also clear that Beauty, during her lucid dream, is not sure whether she will be awakened today.

## 3.2 'Bayesian Updating on Self-locating Evidence Is Impossible'

As mentioned in section 2.1, Pust (2012) has argued that Bayesian updating on self-locating evidence is never possible. The position that this is based on I have dubbed *extreme indexicalism*. I will not use any space arguing against extreme indexicalism, since even if extreme indexicalism is true, Bayesian updating can be saved.

To recap, extreme indexicalism is the position that a temporally indexical proposition can only be entertained at a single moment in time. For example, the sentence "Today is Monday" entertained at 13:00 refers to the proposition $\text{MON}_{13:00}$. According to extreme indexicalism, the proposition $\text{MON}_{13:00}$ is only accessible at 13:00, and thus cannot be assigned a credence at 13:01, or any other moment in time.

To avoid catastrophic consequences, extreme indexicalists need some way to transfer temporal evidence to later moments in time. If an agent knows at 13:00 that it is Monday, she ought to know at 13:01 that it is Monday. Moreover, probabilities of temporal knowledge and the probabilities of conjunctions of temporal knowledge and other knowledge should not normally change as time passes. For example, if the local grocery store is closed on Monday, an agent should not change her credence that the store is closed today between 13:00 and 13:01 as a result of losing

access to $\text{MON}_{13:00}$. Hence, the extreme indexicalist has to accept some sort of continuity principle like the following.

> **Continuity Principle (CP).** Suppose that a rational agent's credence domain at $t_1$ and $t_2$ is generated by a set of sentences $S$, that is, we have $\Omega_t = \{\text{S}_t \mid \text{S} \in S\}$, for $t \in \{t_1, t_2\}$. Suppose that insufficient time has passed to alter the truth value and evidential relevance of whatever propositions the sentences in $S$ signify at $t_1$ and $t_2$. Then her credences are unaltered in the following sense: for all $\text{S} \in S$ we have $P_2(\text{S}_{t_2}) = P_1(\text{S}_{t_1})$.

To recover some types of Bayesian updating, we need only allow that an agent, who receives new indexical evidence at $t_2$, imagines that she had already received it at $t_1$. For example, suppose an agent has a credence function $P_{13:00}$ at 13:00, and learns at 13:01 that it is Monday, that is, she learns $\text{MON}_{13:01}$. To create a new probability function, she first imagines that she learned $\text{MON}_{13:00}$ at 13:00 and used it to create a counterfactual credence function $\tilde{P}_{13:00}$. Clearly, the proposition that MON signifies doesn't change in truth value and evidential relevance between 13:00 and 13:01. Hence, by a counterfactual variant of CP, she should set $P_{13:01}(\cdot) = \tilde{P}_{13:00}(\cdot)$. This reasoning can be summarised in the Counterfactual Continuity Principle.

> **Counterfactual Continuity Principle (CCP).** As before, suppose that a rational agent's credence domain at $t_1$ and $t_2$ is generated by a set of sentences $S$, that is, we have $\Omega_t = \{\text{S}_t \mid \text{S} \in S\}$, for $t \in \{t_1, t_2\}$. Suppose that insufficient time has passed to alter the truth value and evidential relevance of whatever propositions the sentences in $S$ signify at $t_1$ and $t_2$.
>
> Suppose that at $t_2$ the agent learns only $\text{NEW}_{t_2}$ for $\text{NEW} \in S$. Let $\tilde{P}_1$ be her credence function if she knew $\text{NEW}_{t_1}$ at $t_1$. Then her credences at $t_2$ are unaltered with respect to her counterfactual credences at $t_1$, in the following sense: for all $\text{S} \in S$ we have $P_2(\text{S}_{t_2}) = \tilde{P}_1(\text{S}_{t_1})$.

Continuing with the above example, by traditional Bayesian updating, the agent ought to set $\tilde{P}_{13:00}(\cdot) = P_{13:00}(\cdot \mid \text{MON}_{13:00})$. Hence, she ought to set $P_{13:01}(\cdot) = P_{13:00}(\cdot \mid \text{MON}_{13:00})$.

In summary, by combining CCP and traditional Bayesian updating one gets the following updating principle.

> **Extreme Indexicalist Updating (EIU).** As before, suppose that a rational agent's credence domain at $t_1$ and $t_2$ is generated by a set of

sentences $S$, that is, we have $\Omega_t = \{\mathrm{S}_t \mid \mathrm{S} \in S\}$, for $t \in \{t_1, t_2\}$. Suppose a rational agent learns new information $\mathrm{NEW}_{t_2}$ at $t_2$, and suppose that insufficient time has passed to alter the truth value and evidential relevance of whatever propositions the sentences in $S$ signify at $t_1$ and $t_2$. Then she sets $P_2(\cdot) = P_1(\cdot \mid \mathrm{NEW}_{t_1})$.

This principle of updating leads to plausible results. For example, suppose that at 12:00 I am not sure which day it is, and assign each day an equal credence. The grocery store is only closed on Monday, so I have $P_1(\mathrm{OPEN}_{12:00}) = 6/7$. At 12:05, I learn that it is either Monday or Tuesday, that is, $(\mathrm{MON} \vee \mathrm{TUE})_{12:05}$. Applying EIU, I set $P_2(\mathrm{OPEN}_{12:05}) = P_1(\mathrm{OPEN}_{12:00} \mid (\mathrm{MON} \vee \mathrm{TUE})_{12:00}) = 1/2$.

In my argument for thirdism, Beauty uses Bayesian updating after she learns AWAKEN, which states that the researchers awaken her today. Since she has a lucid dream each day, she is sure that it is the same day at 11:00 – during her lucid dream – and after she is awakened at 13:00. Hence, she knows that the truth value and evidential relevance of AWAKEN does not change between 11:00 and 13:00. She can therefore use EIU to set $P_{13:00}(\mathrm{HEADS}) = P_{11:00}(\mathrm{HEADS} \mid \mathrm{AWAKEN}_{11:00})$.

## 3.3 'Bayesian Updating and Elga's Indifference Principle Have Counterintuitive Consequences'

The literature contains many examples of (supposed) counterintuitive consequences that can be derived by combining Elga's restricted principle of indifference (Elga 2004) and various updating principles. Elga's restricted principle of indifference can be used to defend my premise P1 (see section 2.3), and I use Bayesian updating to derive Beauty's post-awakening credences from her lucid dream credences.

In many of these examples, an updating principle is used that is incompatible with type 1 Bayesian updating (as well as extreme indexicalist updating). For example, Meacham's (2008) *many brains argument* uses an updating principle called *centred conditionalisation* to derive that a rational agent should over time become certain that she is a brain in a vat. This argument does not work with type 1 Bayesian updating.[6]

6. The many brains argument involves an agent who at time $t$ learns the proposition "If H is true, there are now $N_t$ brains in a vat with an identical subjective experience as myself," where $H$ is some hypothesis in which the agent has non-zero credence, and $N_t$ increases with time. The truth value of the above sentence clearly changes as time progresses, and therefore, it cannot be conditionalised on with type 1 Bayesian updating or anything similar to it, such as EIU.

Some examples from Bostrom (2007) rely only on a type of updating compatible with type 1 Bayesian updating and Elga's restricted principle of indifference. Hence, it might be argued, we should be suspicious of combining these two principles (as I do), given that they can be used to derive (ostensibly) counterintuitive results.

In the following variation due to Bostrom called *Extreme Sleeping Beauty and Doppelgängers*, Beauty is never woken up after being put to sleep on Monday. Instead, if the coin lands tails, a million doppelgängers of Sleeping Beauty will be created and awoken on consecutive days, starting on Tuesday. These doppelgängers have a subjective experience that is indistinguishable from Beauty's. They will all think they are Beauty and have all of her memories, of which the last one is being put to sleep on Sunday. Hence, after waking up on Monday, Beauty cannot be sure that she is herself or one of the million doppelgängers. A little while after Beauty is woken up on Monday, she is told that she is the real Beauty.

In appendix A.2, I use an argument similar to Bostrom's to show that Beauty's credence in HEADS is 1/1,000,001. The primary assumptions that this argument relies on are Elga's restricted principle of indifference and type 1 Bayesian updating. Elga's restricted principle of indifference requires Beauty to assign equal credence to all locations she might have in a tails-world. There are 1/1,000,002 possible locations, only one of which is the real Beauty's location. Hence, she sets $P(\text{REAL} \mid \text{TAILS}) = 1/1{,}000{,}002$. Beauty uses type 1 Bayesian updating after being told that she is the real Beauty, setting $P_+(\cdot) = P(\cdot \mid \text{REAL})$. The only other assumption this argument uses is $P_+(\text{HEADS}) = 1/2$. (Any other remotely plausible assumptions, such as $P_+(\text{HEADS}) = 2/3$, leads to a similarly counterintuitive result.)

Bostrom claims that it is counterintuitive that Beauty would assign a credence of only 1/1,000,001 to HEADS. Whether this is a problem for my argument, however, is not obvious. The derivation relies not only on type 1 Bayesian updating, but also on a controversial application of Elga's restricted principle of indifference to a science-fiction scenario in which a person doesn't know whether she is herself or one of a million possibly existing doppelgängers – whereas I applied it only to a case of temporal uncertainty. It may simply be replied that it is inappropriate to apply Elga's restricted principle of indifference to cases in which an agent doesn't know who (as opposed to when) she is. One type of response, for example, is that Beauty should remain certain that she is herself. In that case, $P(\text{REAL}) = 1$, and it is easily shown that $P(\text{HEADS}) = 1/2$. But that is no counterargument against thirdism, since the failure of Elga's principle of indifference in *this* case does not imply its failure in a situation in which Beauty is only unsure of the day.

A further problem with this example is that our intuitions are unreliable in such

science-fiction cases. Perhaps some of those who think the result is counterintuitive will reconsider their intuitions based on the following small variation. Suppose now that all million doppelgängers are immediately created after Beauty is put to sleep on Sunday evening. However, they will only be awakened if the coin lands tails. Suppose also that the real Beauty as well as all her doppelgängers have a lucid dream on the morning before their potential awakening.

When Beauty now has her lucid dream on Monday morning, it seems quite reasonable that her credence that she is the real Beauty is 1/1,000,001, since her experience is identical to that of the already existing doppelgängers. Hence, during her dream, she is nearly certain that she is a doppelgänger, and therefore she is nearly certain that she will only be awakened if the coin lands tails. Hence, if she is then awakened, she should be nearly certain that the coin did land tails. In this variant case, the conclusion does not seem counterintuitive.

In the variant case, the conclusion is easier to digest because the doppelgängers' existence is not dependent on the result of the coin toss. In the original, the doppelgängers might not come into existence when the original Beauty wakes up on Monday. One might reason intuitively that on Monday afternoon, she should not assign a credence to being a doppelgänger greater than 1/2, since the doppelgängers will only be created if the coin lands tails. (In fact, thirdist reasoning leads to the contrary conclusion that Beauty's credence in being the real Beauty, after being awakened, is 2/1,000,002.) But this intuitive reasoning is mistaken, since it neglects that Beauty does not know that it is Monday, and therefore, that a million doppelgängers have not already been created. Hence, it does not seem that clear and plausible intuitions are available for this example.

# 4 Halfism Is Untenable

In the previous sections I presented a positive argument for thirdism that I claim is very hard to resist. This does not settle the debate, since, according to some, there exist plausible arguments for both thirdism and halfism. It would take too much space to refute all arguments for halfism. However, all defences of halfism fit into two categories, *double halfism* and *Lewisian halfism*, defined by a central commitment (I explain these positions below). I show that these commitments have unacceptable consequences. Double halfists need to reject type 1 Bayesian updating, which would be an unmotivated departure from traditional Bayesianism. Lewisian halfists need to either reject the principal principle, or take highly implausible positions about the admissibility of Beauty's evidence. Since all current halfists are

either double halfists or Lewisians, this affects all published defences of halfism.

Note that there are other positions in the debate, besides thirdism and halfism, that are unaffected by my arguments. For example, Singer (2014) argues that Beauty's credence in HEADS should be the imprecise interval $[0, 1/2]$. Singer rejects every use of a principle of indifference for self-locating evidence, which both thirdism and halfism rely on.

## 4.1 Double Halfers Must Reject Unproblematic Applications of Bayesian Updating

Double halfism is the position that Beauty should assign a credence of $1/2$ to HEADS both just after she is awakened and after she is told it is Monday. This position relies on rejecting that Beauty can use Bayesian updating on self-locating evidence. As both Elga (2000) and Lewis (2001) argued, if Beauty uses Bayesian updating, she increases her credence in HEADS after being told it is Monday, which for halfists implies $P_+(\text{HEADS}) = 2/3$. Double halfists maintain that this application of Bayesian updating is problematic. This view is attractive because it avoids the counterintuitive conclusion that Beauty has a credence in HEADS greater than $1/2$ after being told it is Monday.

The problem for double halfers is that conditionalisation after Beauty is told it is Monday can be understood as type 1 Bayesian updating. The sentence MON plausibly refers to the same proposition before and after Beauty is told it is Monday, that is, it refers to the proposition $\text{MON}_{\text{Mon}}$. This proposition is accessible both before and after the researchers tell Beauty that it is Monday. Hence, she can use type 1 Bayesian updating and set $P_+(\text{HEADS}) = P(\text{HEADS} \mid \text{MON})$. Extreme indexicalists should also concede that conditionalisation is possible. Suppose 13:00 is a time shortly after Beauty is woken up by the experimenters, and 13:05 is a time after she is told that it is Monday. The truth value and evidential relevance of the propositions signified by MON (that is, $\text{MON}_{13:00}$ and $\text{MON}_{13:05}$) are unchanged, and Beauty knows this. Hence, she can use extreme indexicalist updating to arrive at the same result.

As I argued above, type 1 Bayesian updating is the *prima facie* rational thing to do. Moreover, compelling arguments against type 1 Bayesian updating do not appear to exist. It is possible to apply such updating on self-locating evidence (section 3.2) and Bayesian updating on self-locating evidence does not clearly have counterintuitive consequences (section 3.3).

A third way in which halfers could argue that Beauty shouldn't use Bayesian

updating is by claiming that a different updating rule exists which is to be preferred for updating on self-locating evidence. A popular rule is called the *halfer rule*, of which versions have been defended by Halpern (2004), Meacham (2008) and Briggs (2010). Halfism follows straightforwardly from the halfer rule. However, the halfer rule is controversial (Conitzer 2015; Pittard 2015; Kim 2022). Conitzer shows that the halfer rule can lead to untenable credences (whereas type 1 Bayesian updating leads to the intuitively correct credence). Hence, type 1 Bayesian updating should be preferred in situations in which it is available.

## 4.2 Beauty's Evidence Is Admissible at the Wrong Moments for Lewisian Halfism

According to Lewisian halfists, Beauty uses type 1 Bayesian updating on MON after she is told that it is Monday. If $P(\text{HEADS}) = 1/2$, then, after being told it is Monday, Bayesian updating leads Beauty to set $P_+(\text{HEADS}) = 2/3$.

Lewisian halfists need to make two implausible commitments.

First, they must claim that Beauty's evidence after being told it is Monday is inadmissible for the principal principle (or to reject the principal principle), to avoid $P_+(\text{HEADS}) = 1/2$. Bostrom (2007) argues that our intuitions about admissibility are insufficiently secure to answer this question of admissibility. Bradley (2011) gives an argument that MON is inadmissible.

However, it can be argued that MON is admissible as follows. Consider that MON & CONDITIONS logically entails AWAKEN. Moreover, consider that MON & AWAKEN logically entails CONDITIONS, which can be expressed as $\text{AWAKEN} \leftrightarrow (\text{MON} \vee \text{TUE})$, where $\leftrightarrow$ is the logical biconditional. Hence, Beauty's evidence after learning it is Monday is logically equivalent to {MON, FAIR, AWAKEN}. This is an evidential situation which only involves admissible evidence. Knowledge of the current day is admissible for the outcome of a fair coin toss if the agent has no additional evidence linking these two pieces. Similarly, knowledge that you are awakened today is irrelevant to the outcome of a coin toss in the absence of evidence linking the two. Since CONDITIONS is not part of the evidence {MON, FAIR, AWAKEN}, there is clearly no such evidential link. Hence, Beauty's evidence after learning it is Monday is admissible.

It might be objected that by treating the "if and only if" in CONDITIONS as a material biconditional – instead of a richer modal biconditional – I have left out evidence that Beauty has about a future possibility. That is, in addition to CONDITIONS defined logically, Beauty knows:

> TUECONDITION: On Tuesday, I will be awakened if and only if the coin landed tails.

However, by the principle Logical Irrelevance introduced in section 2.3, the evidence TUECONDITION is irrelevant for TAILS, since on Monday, Beauty is still ignorant with respect to "On Tuesday, I will be awakened." Since it is irrelevant, it is admissible evidence.

Bradley (2011) argues that MON is inadmissible as follows.

> 1. "Today is Tuesday" is inadmissible.
> 2. If an agent with only admissible evidence has two possible pieces of evidence in her credence domain and one piece of evidence is inadmissible, then the other is inadmissible. (404–405)

(By "evidence" in premise 2, Bradley means the agent's evidence prior to learning a piece of the new evidence – in Beauty's case, prior to learning MON.)

Bradley then claims that it follows that MON is inadmissible. I accept both premises and Bradley's arguments for these premises. However, premise 2 is not applicable to Beauty's situation, since before learning MON, she has inadmissible evidence. Hence, it is not the case that Beauty has "only admissible evidence" prior to learning MON.

To see that Beauty had inadmissible evidence just after being awakened and before learning MON, consider that AWAKEN, in combination with her other evidence, implies MON $\vee$ TAILS. As Wallmann and Hawthorne (2020) show, logical disjunctions involving the outcome, with a prior credence less than 1, are inadmissible. The prior credence in MON $\vee$ TAILS before learning AWAKEN is less than one,[7] so AWAKEN is inadmissible.

The admissibility of Beauty's evidence after being told it is Monday is fatal for Lewisian halfism. As Lewis (2001) argued, $P_+(\text{HEADS}) = P(\text{HEADS}) + 1/6$, as follows from Bayesian conditionalisation. Hence, $P_+(\text{HEADS}) = 1/2$, as required by the principal principle, refutes Lewisian halfism.

Second, Lucid Beauty shows that Lewisian halfism has at least one other implausible consequence. Since Lewisian halfists accept type 1 Bayesian updating after Beauty is told it is Monday, they must plausibly also accept type 1 Bayesian updating as used in my argument, when Beauty learns AWAKEN. Given the claim $P(\text{HEADS}) = 1/2$, a Lewisian must reject at least one of my priors P1–P3. First, suppose the Lewisian does not reject $P_*(\text{HEADS}) = 1/2$ nor the independence of

7. Before being awakened by the researchers it is possible that it is Tuesday and the coin landed heads, as can be clearly seen from the perspective of Lucid Beauty's dream.

HEADS and MON. In that case, Beauty must choose a prior probability in MON of 1, that is, she must set $P_*(\text{MON}) = 1$, in order for the Bayesian updating procedure to give $P(\text{HEADS}) = 1/2$. It is clearly irrational for Beauty to be sure it is Monday.

The other strategy would be to deny $P_*(\text{HEADS}) = 1/2$. This, in turn, requires the Lewisian to claim that Beauty's evidence during her lucid dream is inadmissible for the principal principle. However, as I argued in section 2.3, Beauty's evidence during her lucid dream is irrelevant for HEADS, and therefore admissible. In summary, given that the Lewisian accepts Lucid Beauty's use of Bayesian updating, she must accept either of two very implausible consequences.

# 5 Conclusion

This article offered a new argument for thirdism and a set of arguments against all published defences of halfism.

My positive case for thirdism is a new argument using the analogy of Lucid Beauty. This argument works well because its premises are both highly intuitive and easy to defend. The main lines of attack from halfers against this argument use (a) a denial that AWAKEN is new information and (b) a claim that Bayesian updating on self-locating evidence is always inappropriate. I have argued that both approaches fail.

In my negative case for thirdism, I attack the two varieties of halfism directly. Double halfists need to claim that Beauty cannot use seemingly unproblematic applications of Bayesian updating. Lewisian halfists need to claim that Beauty's evidence is inadmissible for the principal principle after being told it is Monday. However, I argued that the admissibility of this evidence is highly plausible, and that existing arguments for the inadmissibility fail. Moreover, Lewisian halfists must reject one of the plausible premises P1–P3 of my positive argument for thirdism, which is another bitter pill to swallow.

## 5.1 Are Halfers' Intuitions to Be Trusted?

If my arguments are correct, it does not look good for halfers. First, my argument for thirdism only seems to rely on plausible premises that most authors, including halfers, would accept. Second, the two variants of halfism, double halfism and Lewisian halfism, turn out to be untenable.

Could there still be a way for halfism to survive? That is a question that may only be answerable by its adherents. However, I suggest that they should not want

to save halfism, since the intuitions that led them there have turned out to be mistaken.

It is crucial for understanding the Sleeping Beauty problem that the proposition AWAKEN is conditional on other facts that are unknown to Beauty prior to learning AWAKEN. That is, AWAKEN will happen if and only if it is Monday or it is Tuesday and the coin landed tails. If someone disregards this conditionality, three intuitive but mistaken lines of reasoning open up leading to halfism.

First, one could reason that AWAKEN is admissible evidence, allowing one to use the principal principle on Monday afternoon. However, AWAKEN is in fact inadmissible, as I argued in section 4.2. Hence, Beauty cannot use the principal principle after being awakened by the researchers.

Second, one could reason that Beauty's relevant evidence on Monday afternoon hasn't changed since Sunday evening, requiring Beauty to retain her Sunday evening credence in HEADS. However, as I argued, AWAKEN, which is equivalent to MON $\vee$ TAILS, is relevant evidence not available on Sunday.

Third, one could reason that Beauty's relevant evidence for HEADS doesn't change before and after being told that it is Monday on Monday afternoon, suggesting double halfism. From this it would follow that Beauty's credence in HEADS before and after learning MON should be the same. (After learning MON, Beauty's credence in HEADS is plausibly 1/2, although Lewis disagreed.) However, before learning MON, Beauty's evidence AWAKEN implies MON $\vee$ TAILS – evidence which favours TAILS, but which is neutralised after learning MON. Hence, MON is new relevant evidence for HEADS.

# A Calculations

## A.1 Lucid Beauty

Let $P_*$ be Beauty's credence function during her lucid dream, when she has the evidence LUCID, MON $\vee$ TUE, FAIR, and CONDITIONS. Then her credence in HEADS conditional on AWAKEN is as follows.

$$P_*(\text{HEADS} \mid \text{AWAKEN}) = \frac{P_*(\text{HEADS \& AWAKEN})}{P_*(\text{AWAKEN})} \tag{1}$$

$$= \frac{P_*(\text{HEADS}) \cdot P_*(\text{MON})}{P_*(\text{AWAKEN})} \tag{2}$$

$$= \frac{1/4}{P_*(\text{AWAKEN})} \tag{3}$$

$$= \frac{1/4}{\begin{array}{l} P_*(\text{AWAKEN} \mid \text{MON}) \cdot P_*(\text{MON}) \\ \quad + P_*(\text{AWAKEN} \mid \text{TUE}) \cdot P_*(\text{TUE}) \end{array}} \tag{4}$$

$$= \frac{1/4}{1 \cdot (1/2) + P_*(\text{TAILS} \mid \text{TUE}) \cdot 1/2} \tag{5}$$

$$= \frac{1/4}{1 \cdot (1/2) + 1/2 \cdot 1/2} = 1/3. \tag{6}$$

In the above calculations, the following principles and priors are used. (1): the definition of conditional probability. (2): logical equivalence of AWAKEN and HEADS & MON given CONDITIONS, prior P3. (3): priors P1 and P2. (4): law of total probability. (5): P1, logical equivalence of AWAKEN and TAILS given TUE. (6): P3, P2.

## A.2 Extreme sleeping Beauty and doppelgängers

Let $P$ be Beauty's credence function after she is awakened by the experimenters. At this point, her evidence consists of:

| | |
|---|---|
| REAL ∨ DOPPEL: | I am either Beauty (REAL) or one of a million doppelgängers (DOPPEL). |
| FAIR: | A fair coin was tossed. |
| CONDITIONS: | The researchers awaken me today if and only if either (REAL) I am Beauty or (TAILS) the coin landed tails. |
| AWAKEN: | The researchers awaken me today. |

After Beauty is awakened, the researchers tell her she is the real Beauty (REAL). Her credence function then becomes $P_+$. We assume $P_+(\text{HEADS}) = 1/2$. After learning REAL, all propositions in her credence domain remain accessible, since no doppelgängers are created in the period before Beauty learns REAL and after she is awakened. Hence, by type 1 Bayesian updating for self-locating evidence, we have

$$P(\text{HEADS} \mid \text{REAL}) = 1/2. \tag{7}$$

Let $p = P(\text{HEADS})$. By Elga's restricted principle of indifference, we have $P(\text{REAL} \mid \text{TAILS}) = 1/1{,}000{,}001$ and $P(\text{REAL} \mid \text{HEADS}) = 1$. By the law of total probability,

$$\begin{aligned} P(\text{REAL}) &= P(\text{REAL} \mid \text{HEADS}) \cdot p + P(\text{REAL} \mid \text{TAILS}) \cdot (1 - p) \\ &= p + \frac{1 - p}{1{,}000{,}001}. \end{aligned} \tag{8}$$

Then, by Bayes' law and (8),

$$\begin{aligned} P(\text{HEADS} \mid \text{REAL}) &= \frac{P(\text{REAL} \mid \text{HEADS}) \cdot p}{P(\text{REAL})} \\ &= \frac{p}{p + (1 - p)/1{,}000{,}001}. \end{aligned} \tag{9}$$

Setting (9) equal to (7), we get

$$\frac{p}{p + (1 - p)/1{,}000{,}001} = 1/2.$$

Solving for $p$ yields $p = 1/1{,}000{,}002$.

# References


Arntzenius, Frank. 2003. 'Some problems for conditionalization and reflection'. *The Journal of Philosophy* 100 (7): 356–370. https://doi.org/10.5840/jphil2003100729.

Bostrom, Nick. 2007. 'Sleeping Beauty and self-location: A hybrid model'. *Synthese* 157 (1): 59–78. https://doi.org/10.1007/s11229-006-9010-7.

Bradley, Darren J. 2011. 'Self-location is no problem for conditionalization'. *Synthese* 182 (3): 393–411. https://doi.org/10.1007/s11229-010-9748-9.

———. 2012. 'Four Problems About Self-Locating Belief'. *The Philosophical Review* 121, no. 2 (April): 149–177. https://doi.org/10.1215/00318108-1539071.

Briggs, Rachael. 2010. 'Putting a Value on Beauty'. In *Oxford Studies in Epistemology,* ed. by Tamar Szabó Gendler and John Hawthorne, 3:3–34. Oxford: Oxford University Press.

Cisewski, Jessi, Joseph B. Kadane, Mark J. Schervish, Teddy Seidenfeld and Rafael Stern. 2016. 'Sleeping Beauty's Credences'. *Philosophy of Science* 83 (3): 324–347. https://doi.org/10.1086/685741.

Conitzer, Vincent. 2015. 'A devastating example for the Halfer Rule'. *Philosophical Studies* 172 (8): 1985–1992. https://doi.org/10.1007/s11098-014-0384-y.

Dorr, Cian. 2002. 'Sleeping Beauty: in defence of Elga'. *Analysis* 62, no. 4 (October): 292–296. https://doi.org/10.1093/analys/62.4.292.

Elga, Adam. 2000. 'Self-locating belief and the Sleeping Beauty problem'. *Analysis* 60, no. 2 (April): 143–147. https://doi.org/10.1093/analys/60.2.143.

———. 2004. 'Defeating Dr. Evil with Self-Locating Belief'. *Philosophy and Phenomenological Research* 69 (2): 383–396. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x.

Halpern, Joseph Y. 2004. 'Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems'. In *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning,* 12–22. AAAI Press.

Horgan, Terry. 2004. 'Sleeping Beauty awakened: new odds at the dawn of the new day'. *Analysis* 64, no. 1 (January): 10–21. https://doi.org/10.1093/analys/64.1.10.

———. 2007. 'Synchronic Bayesian updating and the generalized Sleeping Beauty problem'. *Analysis* 67, no. 1 (January): 50–59. ISSN: 0003-2638. https://doi.org/10.1093/analys/67.1.50.

———. 2008. 'Synchronic Bayesian updating and the Sleeping Beauty problem: reply to Pust'. *Synthese* 160 (2): 155–159. https://doi.org/10.1007/s11229-006-9121-1.

Kim, Namjoong. 2022. 'Sleeping beauty and the evidential centered principle'. *Erkenntnis,* 1–23. https://doi.org/10.1007/s10670-022-00619-6.

Lewis, David. 2001. 'Sleeping Beauty: reply to Elga'. *Analysis* 61, no. 3 (July): 171–176. https://doi.org/10.1093/analys/61.3.171.

Meacham, Christopher J.G. 2008. 'Sleeping beauty and the dynamics of de se beliefs'. *Philosophical Studies* 138 (2): 245–269. https://doi.org/10.1007/s11098-006-9036-1.

Pittard, John. 2015. 'When Beauties Disagree: Why Halfers Should Affirm Robust Perspectivalism'. In *Oxford Studies in Epistemology,* ed. by Tamar Szabó Gendler and John Hawthorne, 5:195–204. Oxford: Oxford University Press.

Pust, Joel. 2008. 'Horgan on sleeping beauty'. *Synthese* 160 (1): 97–101. https://doi.org/10.1007/s11229-006-9102-4.

———. 2012. 'Conditionalization and essentially indexical credence'. *The Journal of philosophy* 109 (4): 295–315. https://doi.org/10.5840/jphil2012109411.

———. 2013. 'Sleeping Beauty, evidential support and indexical knowledge: reply to Horgan'. *Synthese* 190 (9): 1489–1501. https://doi.org/10.1007/s11229-011-9888-6.

Schulz, Moritz. 2010. 'The dynamics of indexical belief'. *Erkenntnis* 72 (3): 337–351. https://doi.org/10.1007/s10670-010-9209-3.

Schwarz, Wolfgang. 2012. 'Changing minds in a changing world'. *Philosophical Studies* 159 (2): 219–239. https://doi.org/10.1007/s11098-011-9699-0.

Singer, Daniel Jeremy. 2014. 'Sleeping beauty should be imprecise'. *Synthese* 191 (14): 3159–3172. https://doi.org/10.1007/s11229-014-0429-y.

Titelbaum, Michael G. 2008. 'The Relevance of Self-Locating Beliefs'. *The Philosophical Review* 117, no. 4 (October): 555–606. https://doi.org/10.1215/00318108-2008-016.

Wallmann, Christian and James Hawthorne. 2020. 'Admissibility troubles for Bayesian direct inference principles'. *Erkenntnis* 85 (4): 957–993. https://doi.org/10.1007/s10670-018-0070-0.