Sleeping Beauty and the Demands of Non-ideal Rationality*

Wolfgang Schwarz

16 January 2025

Abstract. If an agent can't live up to the demands of ideal rationality, fallback norms come into play that take into account the agent's limitations. A familiar human limitation is our tendency to lose information. How should we compensate for this tendency? The Seeping Beauty problem allows us to isolate this question, without the confounding influence of other human limitations. If the coin lands tails, Beauty can't preserve whatever information she has received on Monday: she is bound to violate the norms of ideal diachronic rationality. The considerations that support these norms, however, can still be used. I investigate how Beauty should update her beliefs so as to maximize the expected accuracy of her new beliefs. The investigation draws attention to important but neglected questions about the connection between rational belief and evidential support, about the status of ideal and non-ideal norms, about the dependence of epistemic norms on descriptive facts, and about the precise formulation of expected accuracy measures. It also sheds light on the puzzle of higher-order evidence.

1 Introduction

Bayesian rationality is demanding. The ideal agents of classical Bayesianism always maximize expected utility; they are certain of all logical truths; they never overlook a possible explanation; they never forget anything. Real people are not like that. We do not, and cannot, live up to the demands of ideal Bayesian rationality.

A comprehensive normative theory should tell us not only what ideal agents should do under ideal circumstances, but also what an agent should do if they aren't ideal or if they find themselves in non-ideal circumstances. If you are like me, you should double-check complex proofs; you should write down information that you don't want to lose; you should actively explore alternative explanations.

* For helpful comments, I thank two anonymous reviewers and the participants of a workshop at the Australian National University in 2009, where this material was first presented.

1 That the norms for non-ideal cases may differ from the norms for ideal cases is widely recognized in practical philosophy,

You should do these things even though an ideal agent would never do them.[1]

There have been some attempts within Bayesian epistemology to develop systematic models of non-ideal rationality.[2] The task is somewhat ill-defined, as there are many ways in which one might fall short of ideal rationality. Different shortcomings call for different revisions to the classical Bayesian model.

It can be instructive to consider agents who are imperfect only in a single respect. Along these lines, [Gallow 2021], and [Isaacs and Russell 2023] consider otherwise ideal agents with imperfect discriminatory powers. They argue (convincingly, to my mind) that such agents should not update their beliefs by conditionalization. This is a useful result, if only because it suggest that we, too, should deviate from conditionalization in the relevant respect, given that our many flaws include imperfect discriminatory powers.

I will carry out another controlled study of this type, with a similar methodology. We are going to look at a scenario in which an otherwise ideal agent is in danger of losing information. The scenario in question is the well-known Sleeping Beauty problem.

> *Sleeping Beauty*. On Sunday, Beauty learns of the following arrangement. A fair coin is going to be tossed. If it lands tails, Beauty will be awoken on Monday and on Tuesday, but any traces that Monday will have left on her will be erased, so that she wakes up on Tuesday in the same state in which she woke up on Monday. If the coin lands heads, she will instead be woken up on Monday and made to sleep all through Tuesday.

How should Beauty's beliefs evolve through the course of the experiment? To systematically approach this question, we need a standard by which we can compare candidate belief updates. A popular standard, which I'm going to adopt, evaluates an update by the expected accuracy of the resulting beliefs. It is not entirely obvious how this standard applies to cases involving memory loss and self-locating beliefs. I will argue (in sections 2–5) that the correct application supports "halfing": if Beauty updates her beliefs in the optimal way, she will wake up on Monday with credence ½ in *Heads*.

There are, however, powerful arguments that the epistemic probability of *Heads*, in light of Beauty's Monday morning evidence, is ⅓, not ½. I agree with these arguments. In cases like *Sleeping Beauty*, the norms of non-ideal diachronic rationality clash with the "evidentialist" doctrine that one should proportion one's beliefs to the present evidence. I suggest (in section 6) that this may explain some of the long-standing disagreement over *Sleeping Beauty*.

In section 7, I will suggest that the Sleeping Beauty problem can also shed light on the puzzle of higher-order evidence: If the coin lands heads, Beauty has misleading evidence that her cognitive capacities may be impaired. This "higher-order" evidence should affect her "first-order" beliefs about the coin toss and her location in time.

But first things first.

---

where it is often associated with the "general theory of the second best" of [Lipsey and Lancaster 1956]. See [DiPaolo 2019] for relevant background and arguments that epistemology also needs a "theory of the second best".

2 See, for example, [Hacking 1967], [Earman 1992], [Zynda 1996], [Staffel 2019], and [Skipper and Bjerring 2022].

2 Diachronic rationality

Standard Bayesianism says that when an agent receives new information E, their beliefs should change by (classical) conditionalization, so that the agent's new credence Cr_2 in any proposition A equals their old credence Cr_1 in A conditional on E:

(CC) $\quad \mathrm{Cr}_2(A) = \mathrm{Cr}_1(A/E).$

Many arguments have been given in support of this norm. The argument on which I will focus goes back to [Oddie 1994] and [Greaves and Wallace 2006].

We want to compare different ways in which an agent's beliefs might evolve. Above I spoke of different "updates". Since the new belief state may depend on the world – for example, on which information the agent receives – it is better to think in terms of *update protocols* or *update dispositions*. An update disposition (or protocol) determines a new credence function depending on the old credence function and the state of the world. We can represent such a disposition (or protocol) by a function μ that maps an old credence function Cr_1 and a state of the world w to a new credence function Cr_2. Classical conditionalization is represented by the function that maps Cr_1 and w to $\mathrm{Cr}_1(\cdot/E_w)$, where E_w is the total information the agent receives in state w.

In what follows, we'll mostly be interested in how a fixed belief state should be updated. We can then omit the first argument of μ. The ways in which Cr_1 might change can be represented by functions μ that map a state of the world w to a new credence function Cr_2.

Some functions of this type don't represent realistic update dispositions. If two states w and w' only differ in respects to which the agent's cognitive system is insensitive, the agent can't have a disposition (or follow a protocol) that yields different new beliefs depending on which of w or w' obtains. We'll say that such dispositions (or protocols) are not *available*.

More concretely, let's assume that the agent is about to learn the answer to a certain question, and that the possible answers divide the states into a finite set of propositions $E_1, \ldots, E_n$, which we'll call the *evidence partition*. An update disposition is available iff it doesn't discriminate between states that belong to the same cell of the evidence partition.[3]

Now that we have a formal representation of the available update dispositions, we need a standard by which we can compare them. To this end, we are going to look at the accuracy of the beliefs that a disposition might bring about, where *accuracy* measures how close a belief system comes to the truth.

Formally, an accuracy measure is a function V that assigns a number to a pair of a credence function and a state of the world. Most authors agree that an adequate measure should be "strictly proper".[4] Let's assume that V is some such measure.

We could evaluate an update disposition μ for an agent in a state w by how close the beliefs that would result from μ in w would come to the truth – that is, by $V(\mu(w), w)$. Call this the *actual*

3 Here we assume that the agent has perfect discriminatory powers. See [Gallow 2021] and [Isaacs and Russell 2023].

4 An accuracy measure V is strictly proper if $\sum_w P(w)V(P', w)$ is uniquely maximized for $P' = P$ whenever P is a probability measure. See [Oddie 1997], [Joyce 2009], [Pettigrew 2016], and [Levinstein 2017] in support of strict propriety.

accuracy measure for evaluating update dispositions. Since the most accurate credence function, relative to any state w, is the probability function that assigns 1 to w, the optimal update disposition by the actual accuracy measure would make the agent certain of the true state w: it would make the agent omniscient.

Such dispositions are available. Consider, for example, a disposition that would make the agent absolutely certain of w no matter what they learn and no matter what state they are in. This disposition is available, as it doesn't yield different new beliefs depending on which state within any cell of the evidence partition obtains. An agent with this disposition would have a dogmatic propensity to become completely opinionated, despite their limited evidence. By luck, they would end up with perfectly accurate beliefs. If they had been in any other state (while retaining their disposition), they would have ended up with highly inaccurate beliefs. Such an agent does not seem rational.

I'll return to this intuition in section 4. For now, let's simply accept it. It suggests that we shouldn't just look at the accuracy of the new credence function relative to the actual state of the world. We also need to consider non-actual states. But we shouldn't treat all states equally. If the agent has reason to believe that they are not in a certain state, they may well retain their low credence in that state, even though their new beliefs would be inaccurate relative to that state.

Generalizing, a natural idea is to evaluate an update disposition μ by the average accuracy of its output across all states, weighted by the agent's credence in the relevant state:[5]

$$EV(\mu) = \sum_{w} Cr_1(w) \cdot V(\mu(w), w).$$

This kind of weighted average is called an "expectation". $EV(\mu)$ is the "Cr_1-expected accuracy" of the result of applying μ. Greaves and Wallace [2006] show that the update disposition that maximizes Cr_1-expected accuracy, among all available dispositions, is the conditionalizing disposition that maps every state w to $\mathrm{Cr}_1(\cdot/E_w)$.

This mathematical fact obviously doesn't establish the normative claim that anyone should update their beliefs by conditionalization. One might question the modelling assumptions on which the result is based – for example, the demarcation of available dispositions, or the implicit assumption that the choice of update disposition makes no difference to the state of the world.[6] More simply, one might deny that there are any rational constraints on how beliefs should evolve.

These are important objections, but I want to set them aside. I will use the expected accuracy standard as a proof of concept for how to approach the study of non-ideal rationality. When faced with a non-ideal agent (or an otherwise ideal agent in non-ideal circumstances), we need some way of ordering the available beliefs and update dispositions, so that we can tell which of them are better and which are worse, even if none are ideal. The expected accuracy measure yields such an ordering. I will assume that it gives us at least a clue as to what the norms of non-ideal rationality require, at least in the kinds of cases we'll be looking at.

If we try to use the above measure for Sleeping Beauty, however, we run into a problem. When Beauty wakes up on Monday, she may be unsure whether it is Monday or Tuesday. This uncertainty

5 I assume, here and throughout, that the number of states is finite.

6 On this last point, see, for example, [Greaves 2013], [Schoenfield 2018], or [Konek and Levinstein 2019].

does not pertain to an objective question about the universe as a whole. She is unsure about her present location within the universe. A simple and popular way of representing this kind of uncertainty, due to [Lewis 1979], assumes that credence functions are defined not only over "uncentred" propositions about the universe but also over "centred" propositions whose truth-value might change over time. Let's assume that the states over which credence functions are defined are centred, so that it can be Monday in one state and Tuesday in another. Let μ represent a disposition or protocol for an agent to update their beliefs in the transition from one time to another, from t_1 to t_2. When we evaluate μ, we must take into account that the outputs represent possible beliefs at t_2. We want to evaluate an output Cr_2 by its accuracy not with respect to the state of the world at the earlier time of Cr_1, but with respect to the state of the world at t_2. That is, we should evaluate update dispositions by their expected *future* accuracy, by their accuracy with respect to the state of the world *after the update*.

To model this formally, let's assume that each state (or "centred world") w doesn't just specify what is the case now, but also what is the case at other times. We might, for example, represent a state as a triple (u, i, t) of an uncentred world u, an individual i, and a time t. Suppose the individual i at t in u is about to update their beliefs. There will then be another triple (u, i, t') representing the state when the update is complete. Let's call this state the *doxastic successor* of the original state. Assuming that each state w to which Cr_1 assigns positive credence has a unique doxastic successor $s(w)$, we can express the Cr_1-expected future accuracy associated with an update μ as

$$EV(\mu) = \sum_w Cr_1(w) \cdot V(\mu(s(w)), s(w)).$$

By this revised standard, the optimal update disposition no longer conforms to classical conditionalization, but to a rule that is sometimes called "shifted conditionalization". It says that when an agent receives total evidence E at t_2 then their new credence Cr_2 in any proposition A should equal their

7 See [Schwarz 2017] and references therein for discussion of (SC). To see why updating by (SC) maximizes expected future accuracy, define Cr_1^+ so that $\mathrm{Cr}_1^+(s(w)) = \mathrm{Cr}_1(w)$ for any state w (and $\mathrm{Cr}_1^+(w) = 0$ for any w that is not a successor of any state). Assume that the successor relation is functional and injective. Then Cr_1^+ is a probability measure whenever Cr_1 is. Since $\mathrm{Cr}_1^+(s(w)) = \mathrm{Cr}_1(w)$ for any w, we have $EV(\mu) = \sum_w \mathrm{Cr}_1^+(s(w))V(\mu(s(w)), s(w))$. Summing over $s(w)$ instead of w, and letting E range over the evidence partition, this yields

$$\begin{aligned} EV(\mu) &= \sum_w \mathrm{Cr}_1^+(w)V(\mu(w), w) \\ &= \sum_E \sum_{w\in E} \mathrm{Cr}_1^+(w)V(\mu(w), w) \\ &= \sum_E \sum_{w\in E} \mathrm{Cr}_1^+(w \wedge E)V(\mu(w), w) \\ &= \sum_E \mathrm{Cr}_1^+(E) \sum_{w\in E} \mathrm{Cr}_1^+(w/E)V(\mu(w), w). \end{aligned}$$

As $\mu(w)$ is constant for all $w \in E$, we can write $\mu(w)$ as $\mu(E)$:

$$EV(\mu) = \sum_E \mathrm{Cr}_1^+(E) \sum_{w\in E} \mathrm{Cr}_1^+(w/E)V(\mu(E), w).$$

If V is strictly proper, $\sum_{w\in E} \mathrm{Cr}_1^+(w/E)V(\mu(E), w)$ is maximized by $\mu(E) = \mathrm{Cr}_1^+(\cdot/E)$ (using $P = \mathrm{Cr}_1^+(\cdot/E)$ in the schema from footnote 4). As we can maximise a sum by maximising all its terms, it follows that $EV(\mu)$ is maximized by any function μ that maps each evidence E to $\mathrm{Cr}_1^+(\cdot/E)$. Since $\mathrm{Cr}_1^+(A) = \mathrm{Cr}_1(\{w : s(w) \in A\} = \mathrm{Cr}_1(A^+)$, it follows that $EV(\mu)$ is maximized by any function μ that maps each evidence E to a function Cr_2 such that $\mathrm{Cr}_2(A) = \mathrm{Cr}_1(A^+/E^+)$.

old conditional credence Cr_1 that *A will be the case after the update*, conditional on the assumption that *E will be the case after the update*. Letting A^+ express that A will be the case after the update (formally, $A^+ = \{w : s(w) \in A\}$), the new rule can be expressed as follows:[7]

(SC) $\quad \mathrm{Cr}_2(A) = \mathrm{Cr}_1(A^+/E^+).$

Let's see what this means for Sleeping Beauty.

3 Steadfast halfing

We want to know how Beauty's beliefs should evolve in the transition from Sunday evening, when she goes to sleep, to Monday morning, when she wakes up. It is uncontroversial that her Sunday credence in *Heads* should be ½. Let's assume that it is. We don't know in any detail what information she receives when she wakes up. Does she hear the hum of an air conditioner? Is the lighting dim or bright? Whatever she learns, however, is plausibly irrelevant to *Heads* vs *Tails* from the perspective of her Sunday beliefs. By this I mean that Beauty's Sunday credence in *Heads*, conditional on the hypothesis that these things will be the case when she wakes up, is still ½. Whatever total information E Beauty receives upon awakening, we have[8]

$$Cr_1(Heads^+/E^+) = 1/2.$$

If Beauty's beliefs evolve in line with (SC), it follows that her new credence in *Heads* is ½.

We get a form of halfing. But we don't get "Lewisian halfing", as defended in [Lewis 2001]. Nor do we get "double halfing", as defended, for example, in [Bostrom 2007] or [Meacham 2008].[9] Instead, we get a form of halfing that I will call *steadfast halfing*. According to steadfast halfing, Beauty should give credence ½ to *Heads*, and she should be certain that it is Monday.[10] After all, Beauty is sure on Sunday that it will be Monday when she wakes up:

$$Cr_1(Monday^+/E^+) = 1.$$

By (SC), it follows that $\mathrm{Cr}_2(Monday) = 1$.

We don't need to invoke any general update rule to reach this result. We can show directly that steadfast halfing maximizes expected future accuracy in the transition from Sunday to Monday. To simplify the argument, let's assume that Beauty receives no unexpected evidence on Monday morning: whatever she observes, she already knew on Sunday night that she was going to observe it. Instead of comparing dispositions or protocols for how to respond to different kinds of evidence, we can then simply compare future credence functions.

8 '$Heads^+$' means that it will be true after the update that the coin has landed heads. The time shift isn't doing any interesting work here.

9 Lewisian halfers and double halfers disagree about what Beauty should believe if, later on Monday, she learns that it is Monday. Lewisian halfers say her credence in *Heads* should increase to ⅔, double-halfers say it should stay at ½.

10 Steadfast halfing is defended in [Hawley 2013] on the grounds that it follows from a plausible principle of doxastic "inertia" together with the intuition that Beauty's credence in *Heads* after learning that it is Monday should be ½. I

Let Cr_2 be a candidate credence function for Monday morning. We want to determine the Cr_1-expected accuracy of Cr_2 relative to the state of the world on Monday morning (after the update). We can focus on two questions about that state: whether it is Monday or Tuesday, and whether the coin has landed heads or tails. (Considering further questions would make no difference.) These give us four coarse-grained states: w^{H}_{Mon}, w^{T}_{Mon}, w^{H}_{Tue}, and w^{T}_{Tue}. On Sunday, Beauty's credence is evenly divided between two different states, w^{H}_{Sun} and w^{T}_{Sun}, that have w^{H}_{Mon} and w^{T}_{Mon} as their (respective) successors. The Cr_1-expected future accuracy of Cr_2 is therefore

$$1/2 \cdot V(Cr_2, w^{H}_{Mon}) + 1/2 \cdot V(Cr_2, w^{T}_{Mon}).$$

If V is strictly proper, any function Cr_2 that maximises this expectation gives probability ½ to w^{H}_{Mon} and ½ to w^{T}_{Mon}.[11] It assigns probability ½ to *Heads*, and probability 1 to *Monday*.

So, should Beauty be certain that it is Monday, when she wakes up on Monday? This seems wrong. If Beauty is certain on Monday that it is Monday then she will also be certain *on Tuesday* that it is Monday, provided the coin has landed tails. Her Tuesday beliefs would be highly inaccurate. Shouldn't we take this undesirable consequence into account?[12]

To get clear about this, let's briefly consider a simpler case in which the coin has been tossed before the experiment and Beauty knows all along that it landed tails. Her belief state is certain to be reset on Monday night. How should her beliefs evolve through the course of the experiment?

If Beauty were ideally rational, she would conform to (SC) throughout. She would be certain on Monday that it is Monday and on Tuesday that it is Tuesday. This is incompatible with the constraints of the scenario. Beauty can't wake up on Tuesday with different beliefs than on Monday.[13] Which of the *possible* trajectories would be best?

Beauty could be certain that it is Monday on both awakenings. She would then conform to the ideal norm on Monday, and fall dramatically short of it on Tuesday. It would be better, I think, if Beauty were indifferent between Monday and Tuesday on both awakenings. She would fall short of the ideal norm on both occasions, but less dramatically so. The *aggregated* accuracy across both awakenings would be greater.

Notice that any update disposition that Beauty might employ in the transition from Sunday to Monday effectively determines her beliefs both on Monday morning and on Tuesday morning (in our simplified version of the scenario). There is one output credence Cr_2, but it is instantiated twice, once

will explain in section 6 why I reject this intuition.

11 In the propriety schema from footnote 4, let P be the probability function that assigns ½ each to w^{H}_{Mon} w^{T}_{Mon}.

12 One might also object that Beauty can't rule out w^{T}_{Tue}, since her internal state would be just the same if she were in w^{T}_{Tue}. This objection seems to rely on an "evidentialist" principle that I'll discuss (and reject) in section 6. It also relies on some kind of internalism. As [Russell 2023] points out, a certain externalist conception of evidence might imply that Beauty has decisive evidence for *Monday*. See [Weatherson 2015], however, for (externalist) doubts about such a conception.

13 I have stipulated (with [Elga 2000: 143, fn.2] and others) that if the coin lands tails then Beauty's internal state on Tuesday morning is reset to match her internal state on Monday morning. Plausibly, different beliefs about her temporal location would require some difference in her internal state. Some authors merely stipulate that Beauty's Tuesday awakening is "indistinguishable" from her Monday awakening, or that her "memory of Monday" is "erased", or that she has "the same total evidence" on both awakenings. Without some (questionable) premises connecting evidence, memory, and indistinguishability to belief, these stipulations don't settle whether Beauty can conform to (SC) throughout the experiment, even if the coin lands tails. If she can, I'd say that she should.

on Monday and once on Tuesday. Let's say that an agent undergoes *doxastic fission* (at a given state) if there are multiple future points in their doxastic trajectory at which the output of the next update will be instantiated, no matter which update they choose. I suggest that if an agent undergoes doxastic fission then all instances of the output credence should figure equally in the assessment of the update. Call this the *doxastic fission rule*.

Intuitively, if the new credence is certain to be instantiated twice, on Monday and on Tuesday, then both of these locations have equal claim to count as "after the update" in the expected accuracy calculation. This calls for a generalization of the measure introduced in the previous section. We've assumed that every state w is succeeded by a unique state $s(w)$ that obtains after the update has been applied in w. We need to make room for cases in which w has more than one successor. The doxastic fission rule narrows down the possible generalizations. In our simplified Sleeping Beauty scenario, it implies that Beauty should be indifferent between *Monday* and *Tuesday*. The rule does not yet determine an answer to the original Sleeping Beauty problem, where Beauty doesn't know whether she will undergo doxastic fission. Here, we need to decide whether the mere possibility of doxastic fission affects what Beauty should do, even if the coin lands heads. In the next section, I will argue that it does. We then need to decide how to balance the aggregated accuracy of the two Cr_2 instances in case of *Tails* against the accuracy of the single Cr_2 instance in case of *Heads*. In section 5, I will defend an answer that leads to Lewisian halfing.

Before we continue, a few quick comments on the doxastic fission rule.[14] As stated above, the rule covers a very narrow range of cases. Return to the simplified Sleeping Beauty scenario where Beauty's belief state is sure to be reset on Monday night. Suppose she receives some evidence immediately upon awakening on Monday that is different from the evidence she receives on Tuesday. The rule then doesn't apply because the Monday morning credence is different from the Tuesday morning credence. But Beauty's Tuesday credence is still a direct result of the update disposition chosen on Sunday. The rule should plausibly be extended to cover this type of case. It should probably also be extended to cases in which a later belief state is reset to *some function* of an earlier belief state, rather than to the earlier belief state itself. More complex cases might be considered. For example, what if Beauty's credence is reset only if the update from Sunday to Monday leaves her with a high credence in *Monday*? I would like to have a fully general rule. Alas, I have none to offer. I'm inclined to think that an update should be evaluated by the entire future trajectory of the agent's beliefs, as in the "sophisticated choice" approach to dynamic decision theory, but this raises thorny questions that I'm not sure how to resolve.[15] For the Sleeping Beauty problem, the simple doxastic fission rule is all we

14 I thank an anonymous reviewer for prompting me to address the following points.

15 One family of thorny problems arises from "epistemic bribes" (compare [Greaves 2013]): suppose you will lose all your geographical knowledge tonight unless you now become confident that the moon is made of cheese; knowing this, should you become confident that the moon is made of cheese? A different kind of problem arises from uncertainty about the length of one's doxastic trajectory. Suppose a fair coin is tossed tonight: heads you will die soon after waking up; tails you will live a long life. If we assess your belief update by the expected *average* accuracy of your future beliefs, we get the implausible verdict that you should wake up confident that the coin has landed heads. If instead we go by the expected *total* accuracy of your future beliefs, we get the implausible verdict that you should wake up confident that the coin has landed tails, if we add the assumption that you will forget that you would have died if the coin had landed heads, so that your credence in the outcome will remain the same throughout your long life. (In section 5, I will defend an averaging rule for direct doxastic successors. I do *not* advocate averaging over entire future trajectories.)

need.

4 Normative state dependence

Whether Beauty can comply with the ideal norm (SC) depends on the outcome of the coin toss. If the coin lands tails, she is bound to violate (SC) in the transition from Monday to Tuesday. But nobody is going to tinker with her beliefs if it lands heads. In that case, she is made to sleep through all of Tuesday, but this is no impediment to her rationality. One might conclude that fallback norms only come into play if the coin lands tails; if it lands heads, Beauty can, and should, comply with (SC).

The proposal is not that Beauty should have an update disposition that is sensitive to the unknown outcome of the coin toss. Such dispositions are not available. Rather, the proposal is that *which update disposition she ought to have* is sensitive to the outcome of the coin toss. The *norms* would depend on an unknown aspect of the state.

This kind of "state-dependence" of the norms is incompatible with a broadly internalist conception of rationality. Our expected (future) accuracy measure embraces the relevant internalism: the optimal update disposition, in terms of expected (future) accuracy, never depends on external facts about which state is actual.[16] One might argue that this is a problem – that what's rational *can* depend on such facts.

What's rational might, for example, depend on whether the agent is in a skeptical scenario. Suppose you are looking at what appears to be a red wall. One might suggest that you can be rationally confident that the wall is red *if it is in fact red*, but not if it is only made to look red by clever lighting, even though your cognitive system is not sensitive to the difference between the two possibilities.[17] This might be relevant to Sleeping Beauty, for isn't the *Tails* $\wedge$ *Tuesday* scenario a kind of skeptical scenario? (Compare [Hawley 2013].)

The study of bounded rationality might also motivate state-dependence of rational norms. Friends of "ecological rationality" emphasize that how an agent should reason depends on contingent facts about the agent and their environment, since heuristics that work well in one kind of environment fare badly in another. (See, for example, [Todd and Gigerenzer 2012].) If what's rational depends on these facts – and not just on the agent's beliefs about them – then the norms of rationality are state-dependent.

We could make room for state-dependence in our expected accuracy framework by conditionalizing the expectation on the relevant aspect of the state. That is, if X is an aspect of the state on which the norms depend, we could evaluate an update disposition μ *in a state of type* X by

$$EV(\mu/X) = \sum_w Cr_1(w/X)V(\mu(s(w)), s(w)).$$

This is equivalent to taking the weighted average only over worlds where X obtains and renormalizing the weights to factor out the probability of X.

16 There's another respect in which our measure may be externalist. The optimal update disposition depends on what's available, which depends on the partition $\{E_1, \ldots, E_n\}$. I have assumed that this partition is somehow fixed. In reality, the agent's sensitivity to the world may vary from state to state, and it may not be known by the agent.

17 Such a view might be inspired by [Williamson 2000]. Williamson himself does not explicitly subscribe to state-

What's optimal now depends on which aspects of a state the norms are sensitive to. If the norms are sensitive to every aspect, we get the "actual accuracy measure" from section 2: we'd conclude that rationality requires becoming omniscient. We get a similarly absurd result in the Sleeping Beauty problem if we allow the norms of rationality to depend on the outcome of the coin toss. The optimal update in case of heads then doesn't lead to steadfast halfing, but to "steadfast oneing": Beauty would wake up certain that it is Monday *and that the coin has landed heads*.[18] In case of tails, she would similarly have to wake up certain that the coin has landed tails.

Beauty could conform to these demands. She could, for example, have a dogmatic disposition to become certain that the coin has landed heads, no matter what. If the coin does land heads, she would conform to the state-dependent norm. But her accuracy would be sheer luck. This kind of dogmatism seems irrational.

Neither of the above motivations for state-dependence supports the view that the norms may depend on the outcome of an ordinary coin flip. If you know that the wall in front of you has either been painted red or made to look red by clever lighting, depending on the outcome of a coin flip, then you can't rationally be confident that the wall is red, even if in fact it is. Skeptical scenarios with a known positive chance must be given proportionate credence. Likewise, the ecological conception of bounded rationality becomes untenable if different outcomes of a coin flip are treated as different environments: we don't want to say that in an environment where a coin has landed heads people should be confident (without any relevant evidence) that the coin has landed heads, even though this would be a good "heuristic" in that kind of environment.

The general question of normative state-dependence deserves more careful study. The above considerations suggest a constraint on state-dependence – a kind of "anti-luck" condition: The rational norms do not depend on aspects of the environment that could easily have been different. It follows that the norms of non-ideal rationality are not just norms for agents who can't comply with the ideal norms. If the coin lands heads, Beauty *can* comply with the ideal norm (SC), but it would be wrong for her to do so. The norms of non-ideal rationality come into play merely because Beauty can't be sure that she is ideal. Beauty's uncertainty about her ideality seems to *make* her non-ideal, insofar as it makes her violate the ideal norms. (Compare [Christensen 2007].) I will return to this observation in section 7.

5 Lewisian Halfing

Let's recompute how Beauty should update her beliefs in the transition from Sunday to Monday. Unlike in section 3, we'll take into account that the new belief state may be instantiated twice.

As before, we can assume that Beauty receives no unexpected evidence when she wakes up, so that we can focus on the expected accuracy of future credence functions. Let Cr_2 be a candidate successor to Beauty's Sunday credence. If Beauty adopts a disposition that leads to Cr_2, then Cr_2 will either be instantiated once, in $w^{\mathrm{H}}_{\mathrm{Mon}}$, or it will be instantiated twice, in $w^{\mathrm{T}}_{\mathrm{Mon}}$ and in $w^{\mathrm{T}}_{\mathrm{Tue}}$, depending on the

dependence. He rather suggests that you should proportion your beliefs to your "evidence", and that the relevant evidence is different depending on the true colour of the wall.

18 In general, any function μ that maximizes $EV(\mu/X)$ outputs a credence function that is certain of X.

outcome of the coin toss. By the doxastic fission rule, we should give equal consideration to Cr_2's accuracy in $w^{\mathrm{T}}_{\mathrm{Mon}}$ and $w^{\mathrm{T}}_{\mathrm{Tue}}$. But how should we balance these accuracy scores against Cr_2's accuracy in $w^{\mathrm{H}}_{\mathrm{Mon}}$?

I suggest that the aggregated *Heads* and *Tails* scores should be weighed in accordance with their probability. Since Beauty is 50% confident that Cr_2 will be instantiated once, in $w^{\mathrm{H}}_{\mathrm{Mon}}$, the accuracy of Cr_2 in w^{H}_{Mon} should have weight ½ in the assessment of the update; the accuracy of Cr_2 in $w^{\mathrm{T}}_{\mathrm{Mon}}$ and $w^{\mathrm{T}}_{\mathrm{Tue}}$ should each have weight ¼, so that the total weight of the *Tails* instances is also ½. The Cr_1-expected value of updating to Cr_2 is then

$$1/4 \cdot V(Cr_2, w^{\mathrm{T}}_{\mathrm{Mon}}) + 1/4 \cdot V(Cr_2, w^{\mathrm{T}}_{\mathrm{Tue}}) + 1/2 \cdot V(Cr_2, w^{\mathrm{H}}_{\mathrm{Mon}}).$$

If V is strictly proper, any credence function Cr_2 that maximizes this sum assigns probability ½ to *Heads* and ¾ to *Monday*.[19] We get standard (non-steadfast) halfing.

In general, I suggest the following measure to deal with cases of possible doxastic fission. Let's redefine $s(w)$ as the set of w's doxastic successors. I suggest that we assess a candidate update μ by *averaging* the accuracy of multiple successors:

$$EV(\mu) = \sum_w Cr_1(w) \sum_{w' \in s(w)} \frac{V(\mu(w'), w')}{|s(w)|}.$$

Other measures are conceivable. In particular, one might suggest to *sum* over the successors:

$$EV(\mu) = \sum_w Cr_1(w) \sum_{w' \in s(w)} V(\mu(w'), w').$$

This is equivalent to giving each successor the full weight of its predecessor, allowing the sum of the weights to exceed 1. It supports thirding: since the two Cr_2 instances in case of *Tails* would both get weight ½, the Cr_1-expected value of updating to Cr_2 would be

$$1/2 \cdot V(Cr_2, w^{\mathrm{T}}_{\mathrm{Mon}}) + 1/2 \cdot V(Cr_2, w^{\mathrm{T}}_{\mathrm{Tue}}) + 1/2 \cdot V(Cr_2, w^{\mathrm{H}}_{\mathrm{Mon}}).$$

This is maximized by a function Cr_2 that assigns ⅓ to *Heads*.[20]

To decide between these two standards, between the "average measure" and the "total measure", let's look at a situation in which they make a difference even though the agent's rationality is not under threat:[21]

Fred's home planet, *Sunday*, is surrounded by two moons, *Monday* and *Tuesday*. Tonight,

19 In the schema from footnote 4, let P be the probability function that assigns ¼ each to $w^{\mathrm{T}}_{\mathrm{Mon}}$ and $w^{\mathrm{T}}_{\mathrm{Tue}}$ and ½ to $w^{\mathrm{H}}_{\mathrm{Mon}}$.

20 [Kierland and Monton 2005] make a similar observation. They note that thirding maximises the chance-expected total accuracy of Beauty's new credence, while halfing maximises the chance-expected average accuracy. They argue that both assessments of credal accuracy are defensible. Unlike Kierland and Morton, I am evaluating Beauty's new credence from the perspective of her earlier credence, not from the perspective of objective chance. For reasons that should become clear in section 6, I do not believe that Beauty's credence should maximize any kind of chance-expected accuracy.

21 This kind of case has been introduce into the *Sleeping Beauty* debate in [Lewis 2007]. My staging follows [Schwarz 2015].

> while Fred is asleep, he will be teleported to Monday: his body will be scanned and destroyed and recreated from local matter on Monday. Depending on the outcome of a fair coin flip, he will also be teleported to Tuesday – tails he will, heads he won't.

Let's not worry about whether the teleportation products are the same person as the original Fred. Our question is which update would be best in terms of expected future accuracy.[22] Since all teleportation products have the same beliefs upon awakening, the output of any update disposition Fred might adopt is instantiated twice if the coin lands tails and once if it lands heads. In this respect, the case is analogous to *Sleeping Beauty*. By the average measure, the optimal update disposition would retain Fred's credence 1/2 in *Tails*. By the total measure, his credence should increase to 2/3.

Generalizing, the total measure suggests that if you are initially undecided between a hypothesis – say, Everettian Quantum Mechanics – according to which you frequently undergo personal fission and an alternative hypothesis according to which you do not, then you should become increasingly confident in the first hypothesis, even if you receive no relevant (unexpected) evidence. This seems wrong.

We can't use expected accuracy considerations to show that the total measure goes wrong here, since what's at issue is precisely how these considerations should be spelled out. But we can invoke other ideas about diachronic rationality to support the judgement. For example, suppose Fred is ideally rational, and knows that he is. If Fred's credence in *Tails* increases to 2/3 then his original credence in *Tails* on Sunday comes apart from what he knows to be his (or his successors') future credence. He violates an attractive *Reflection* principle, according to which an agent's present credence in uncentred propositions should equal the expectation of their future credence. Arguably, this kind of disagreement with one's future self never happens to ideal agents who know that they are ideal.[23]

The total measure has other problems as well. If the accuracy scores $V(\mathrm{Cr}, w)$ are positive (greater than 0), it supports the repugnant conclusion that it would be better, from a purely epistemic perspective, to fission into many people with highly inaccurate beliefs than to remain unfissioned with highly accurate beliefs. If accuracy scores are negative, the approach leads to the opposite, but equally repugnant, conclusion that it would be better to remain a single person with highly inaccurate beliefs than to fission into lots of people with highly accurate beliefs. These conclusions are not only implausible on their own, it is also odd that the demands of epistemic rationality should depend on whether accuracy scores are positive or negative. The average measure has none of these problems.

You may still want to resist my argument for halfing. Aren't there powerful arguments for thirding? There are indeed. We'll look at some of them in a moment. First, let's check whether the present approach supports "Lewisian halfing" or "double halfing". The two views agree on what Beauty should believe when she wakes up on Monday. They disagree over what she ought to believe if she is later told that it is Monday. Lewisian halfers say that her credence in *Heads* should increase to 2/3; double halfers say that it should remain at 1/2.

Let's assume that Beauty has woken up giving credence 1/2 to $w^{\mathrm{H}}_{\mathrm{Mon}}$ and 1/4 to each of $w^{\mathrm{T}}_{\mathrm{Mon}}$ and

22 I assume that the states in which the teleportation products are located qualify as doxastic successors of Fred's pre-fission state. Pace [Hedden 2015], I don't think this commits me to any view about the metaphysics of personal identity.

23 See [van Fraassen 1984], [Hild 1998], [Easwaran 2013], [Huttegger 2013], and [Gallow 2021], among others, in support of Reflection.

$w^{\mathrm{T}}_{\mathrm{Tue}}$, as both kinds of halfing recommend. The information that it is Monday rules out one of the two *Tails* possibilities and no *Heads* possibility. We might expect it to increase Beauty's credence in *Heads*. We can confirm this by looking at the expected future accuracy of the update. The three states $w^{\mathrm{H}}_{\mathrm{Mon}}$, $w^{\mathrm{T}}_{\mathrm{Mon}}$, and $w^{\mathrm{T}}_{\mathrm{Tue}}$ all have a unique successor, where Beauty learns the day of the week[24] – let's call them $w^{\mathrm{H}}_{\mathrm{Mon2}}$, $w^{\mathrm{T}}_{\mathrm{Mon2}}$, and $w^{\mathrm{T}}_{\mathrm{Tue2}}$, respectively. The choice between the average and the total measure doesn't matter. Either way, we have

$$EV(\mu) = 1/2V(\mu(w^{\mathrm{H}}_{\mathrm{Mon2}}), w^{\mathrm{H}}_{\mathrm{Mon2}}) + 1/4V(\mu(w^{\mathrm{T}}_{\mathrm{Mon2}}), w^{\mathrm{T}}_{\mathrm{Mon2}}) + 1/4V(\mu(w^{\mathrm{T}}_{\mathrm{Tue2}}), w^{\mathrm{T}}_{\mathrm{Tue2}}).$$

As Beauty receives the same information (that it is Monday) at $w^{\mathrm{H}}_{\mathrm{Mon2}}$ and at $w^{\mathrm{T}}_{\mathrm{Mon2}}$, the available update functions return the same credence for these states. Let's write $\mu(\mathrm{Mon})$ for $\mu(w^{\mathrm{H}}_{\mathrm{Mon2}})$ and $\mu(w^{\mathrm{T}}_{\mathrm{Mon2}})$, and $\mu(\mathrm{Tue})$ for $\mu(w^{\mathrm{T}}_{\mathrm{Tue2}})$. We have:

$$\begin{aligned} EV(\mu) &= 1/2V(\mu(Mon), w^{\mathrm{H}}_{\mathrm{Mon2}}) + 1/4V(\mu(Mon), w^{\mathrm{T}}_{\mathrm{Mon2}}) + 1/4V(\mu(Tue), w^{\mathrm{T}}_{\mathrm{Tue2}}) \\ &= 3/4[2/3V(\mu(Mon), w^{\mathrm{H}}_{\mathrm{Mon2}}) + 1/3V(\mu(Mon), w^{\mathrm{T}}_{\mathrm{Mon2}})] + 1/4V(\mu(Tue), w^{\mathrm{T}}_{\mathrm{Tue2}}). \end{aligned}$$

If V is strictly proper, the first term, and thereby the entire sum, is maximized iff $\mu(\mathrm{Mon})$ assigns $2/3$ to $w^{\mathrm{H}}_{\mathrm{Mon2}}$ and $1/3$ to $w^{\mathrm{T}}_{\mathrm{Mon2}}$.[25] We get Lewisian halfing.

One might think that Beauty would prefer a double-halfer disposition if she could already fix her update dispositions for the entire experiment on Sunday.[26] More simply, suppose Beauty immediately learns the day of the week when she wakes up. Given that her Sunday credence in *Heads* is $1/2$, wouldn't she want her credence after awakening to be $1/2$ as well? She would not. Still using 'Mon2' and 'Tue2' as indices for states where the day of the week is revealed, the output of Beauty's Sunday-night update will be located either at $w^{\mathrm{H}}_{\mathrm{Mon2}}$ (if the coin lands heads) or at both $w^{\mathrm{T}}_{\mathrm{Mon2}}$ and $w^{\mathrm{T}}_{\mathrm{Tue2}}$, (if the coin lands tails). By the average future accuracy measure[27],

$$EV(\mu) = 1/2V(\mu(w^{\mathrm{H}}_{\mathrm{Mon2}}), w^{\mathrm{H}}_{\mathrm{Mon2}}) + 1/4V(\mu(w^{\mathrm{T}}_{\mathrm{Mon2}}), w^{\mathrm{T}}_{\mathrm{Mon2}}) + 1/4V(\mu(w^{\mathrm{T}}_{\mathrm{Tue2}}), w^{\mathrm{T}}_{\mathrm{Tue2}}).$$

This is the same expression as above. It is maximized by an update μ that outputs credence $2/3$ in *Heads* when Beauty learns that it is Monday.

Let me remind you how we got here. We started with a standard for evaluating belief updates: the expected (future) accuracy standard. When we tried to apply this to the Sleeping Beauty problem, we encountered some questions: Can the norms of (non-ideal) rationality depend on the outcome of a coin flip? Are Beauty's Tuesday beliefs (if the coin lands tails) relevant to how she ought to update her beliefs on Sunday night? How should states with multiple "successors" be weighed against states with a single successor? I have offered arguments for a certain combination of answers, a combination that supports Lewisian halfing. My real aim, however, is not to defend Lewisian halfing. My aim is to illustrate how considerations of diachronic rationality can systematically apply to a case like

24 I assume Beauty learns on Tuesday that it is Tuesday, if only by noticing that she is not told that it is Monday.

25 In the schema from footnote 4, let P be the probability function that assigns $2/3$ to $w^{\mathrm{H}}_{\mathrm{Mon2}}$ and $1/3$ to $w^{\mathrm{T}}_{\mathrm{Mon2}}$.

26 I thank a reviewer for raising this issue.

27 We here need the extended version of the doxastic fission rule mention at the end of section 3 that covers cases in which the fission products receive different evidence.

Sleeping Beauty, where compliance with the ideal norms may be impossible. If beliefs are subject to diachronic norms, we need to *address* the above questions. The literature on Sleeping Beauty has almost entirely ignored them.[28]

6 Sleeping Beauty: A diagnosis

The Sleeping Beauty problem has divided philosophers for more than 20 years. It is unlikely that one side is simply making a calculation mistake. A more plausibly conjecture is that the disagreement over Sleeping Beauty, like that over Newcomb's Problem, traces back to a more fundamental disagreement about the nature of rationality.

I have asked how Beauty's beliefs should evolve from Sunday to Monday (and Tuesday). Many authors focus on a different question. They ask to what extent Beauty's evidence, when she wakes up on Monday, supports the hypothesis that the coin has landed heads. The answer to *this* question is plausibly 1/3.

Let's go through Beauty's evidence. Beauty remembers the general experimental setup (S), she realizes that she is awake (A), and that she has no memories from later than Sunday (R). By itself, S lends equal support to the four combinations of *Heads* and *Tails* with *Monday* and *Tuesday*. R rules out any further possibilities (such as *Heads* $\wedge$ *Wednesday*); A excludes *Heads* $\wedge$ *Tuesday*. The remaining possibilities therefore have probability 1/3 each. Two of them are *Tails* possibilities. The probability of *Heads*, in light of Beauty's evidence, is therefore 1/3.[29]

We can corroborate this result by first determining the evidential probability of *Heads* conditional on $E \wedge$ *Monday*, where E is Beauty's total evidence upon awakening. As before, Beauty's information S about the scenario confers equal probability to *Heads* and *Tails*. The remainder of $E \wedge$ *Monday* rules out all but one "centre" in each of the uncentred worlds left open by S, without ruling out any of these uncentred worlds. The effect of this purely centred information is plausibly to allocate the probability of each previously open uncentred world to the uniquely open centre within it.[30] So we have $\Pr(\textit{Heads}/E \;\&\; \textit{Monday}) = 1/2$, where 'Pr' expresses evidential probability. We also have $\Pr(\textit{Monday}/E \wedge \textit{Tails}) = 1/2$ and $\Pr(\textit{Monday}/E \wedge \textit{Heads}) = 1$. By Bayes' Theorem, it follows that

28 A rare precedent for the present study are [Arntzenius 2002] and [Arntzenius 2003]. Arntzenius notes that standard considerations of diachronic rationality support steadfast halfing, and that this seems to ignore the threat of cognitive malfunctioning.

29 Variations of this argument can be found in many places, starting with [Piccione and Rubinstein 1997], [Dorr 2002], and [Arntzenius 2003]. The canonical formulation is [Horgan 2004]. I assume, with [Horgan 2008] and against [Pust 2008], that *Heads* $\wedge$ *Tuesday* can have positive evidential probability even though Beauty is not awake if it is Tuesday and the coin lands heads. (I don't see why she would have to be unconscious, as Pust assumes. She could be dreaming. Note that my argument for Lewisian halfing would go through even if she was awake, as long as her memory isn't reset.) There's a structural difference here between *Sleeping Beauty* and the "Fissioning Fred" case from section 5: Fred doesn't even exists on Tuesday if the coin lands heads. *His* evidence arguably confers probability 1/2 to *Heads* when he wakes up on Monday.

30 Here I am using an "anthropic principle". See [Arntzenius and Dorr 2017] and [Isaacs et al. 2022] for what the general principle might look like; all the principles discussed in these papers agree about the present case, provided that Beauty counts as an "observer" while asleep on Tuesday if the coin lands heads. (See the previous footnote.) See also [Pust 2023] for an alternative argument for $\Pr(\textit{Heads}/E \wedge \textit{Monday}) = 1/2$ based on principles of "direct inference".

$\Pr(Heads/E) = 1/3$.[31]

I find these, and other, arguments convincing. To resist them, one would probably have to deny that standard (synchronic) Bayesian reasoning can be applied to self-locating propositions (as suggested, e.g., in [Meacham 2008] and [Builes 2020]). I see no good motivation for this denial, especially as the proposed alternatives have highly counterintuitive implications.[32] Instead, I accept the conclusion: the probability of *Heads*, in light of Beauty's evidence, is 1/3. If epistemic rationality were a matter of proportioning one's beliefs to the present evidence, thirding would be the correct answer to the Sleeping Beauty problem. But I don't think epistemic rationality is simply a matter of proportioning one's beliefs to the present evidence. This, I suspect, is the deeper normative question that lies behind much of the disagreement over Sleeping Beauty.

Note that classical Bayesianism does not endorse the "evidentialist" assumption. If an agent updates their beliefs by conditionalization then their beliefs at a given time depend on their evidence at that time *and on their earlier beliefs* – whether or not the earlier beliefs are part of the present evidence. In effect, classical Bayesianism requires an agent's beliefs to be proportioned to their *past and present* evidence.

Some have found this problematic, as it seems to leave no room for forgetting.[33] I disagree. From an epistemic perspective, retaining information really is better than losing it. An ideally rational agent who optimally updates their beliefs over time – as measured, for example, by our expected accuracy standard – would never lose information.

Real agents, of course, are bound to lose information. A model of *non-ideal* rationality needs to explain how an agent's beliefs should evolve if the agent can't store everything they have ever learned. It should also explain how an agent should deal with a threat of unintended information loss: how they should react to the possibility that they will lose information that they should have retained. Sleeping Beauty finds herself in just this kind of situation. We have seen that the mere threat of information loss is enough to bring non-ideal norms onto the table. Beauty should not update her beliefs by conditionalization or shifted conditionalization. The considerations that support these rules under ideal conditions support a different kind of update – an update that leads to Lewisian halfing.

You may disagree with my claim that losing information is a sign of imperfect rationality. Perhaps you think (with [Moss 2015] and [Hedden 2015]) that present beliefs are never rationally constrained by earlier beliefs. If so, I don't expect to change your mind. (Not in this paper, anyway.) But I hope you can see that our disagreement matters for the Sleeping Beauty problem.

It is not a coincidence that the standard argument for halfing, first discussed in [Elga 2000] and [Lewis 2001], is diachronic. In short, the argument is that Beauty's Sunday credence in *Heads* should be 1/2, and that she doesn't acquire relevant new evidence when she wakes up on Monday. (The argument of the present paper may be seen as a refinement of this line of thought.) Most arguments

31 That Beauty's credence in *Heads* conditional on *Monday* should be 1/2 is related to the popular idea (first stated in [Elga 2000]) that her credence in *Heads* should be 1/2 if she has found out that it is Monday. By focussing on her conditional credence we can discharge any assumptions about how Beauty should respond to the evidence that it is Monday.

32 See, for example, [Bradley 2011: sec.9], [Titelbaum 2013: ch.10], [Button et al. 2024]. Needless to say, I have not *shown* that standard synchronic Bayesian reasoning is applicable to self-locating propositions. Throughout this paper, I am simply taking for granted that it is.

33 See, for example, [Talbott 1991], [Williamson 2000], [Arntzenius 2003].

for thirding, by contrast, rely only on judgments about evidence. Sometimes the evidentialist premise is explicitly stated. Often it is hidden.

Consider, for example, the following argument for thirding, presented in [Piccione and Rubinstein 1997], [Dorr 2002], and [Arntzenius 2003]. The argument starts with a variation of the Sleeping Beauty scenario in which the memory reset takes place no matter how the coin lands. Either way, Beauty is woken up on Monday and on Tuesday, with her memory reset before the second awakening. The only effect of the coin toss is that if the outcome is heads then Beauty receives a special signal soon after waking up on Tuesday. In this scenario, Beauty's credence in *Heads* when she wakes up on Monday should clearly be ½. When she then doesn't receive the signal, she can rule out one of the two previously open *Heads* possibilities: her credence should decrease to ⅓. I agree. The modified scenario is now said to be analogous to the original scenario. But while the two scenarios are analogous with respect to Beauty's evidence about the coin toss, they differ in another respect. In the modified scenario, Beauty knows on Sunday that her updated beliefs will be instantiated twice, no matter how the coin lands. This makes a difference to how Beauty should update her beliefs. The expected accuracy standard recommends thirding in the variant scenario, but halfing in the original.[34]

Admittedly, some authors have offered diachronic arguments for thirding, suggesting that it can be derived from general norms of ideal (!) diachronic rationality. (See, e.g., [Titelbaum 2008], [Kim 2009], [Briggs 2010], [Schulz 2010], [Moss 2012], and [Spohn 2017].) A detailed discussion of these arguments would take us too far afield. We already know from the previous sections that the proposed norms could not be supported by the expected accuracy standard. In some cases, they assume an evidentialist treatment of self-locating beliefs: at any time, the agent is supposed to figure out their temporal location from scratch, based only on their present evidence. I see no good motivation for such a hybrid approach. If the old credence constrains the new credence about eternal matters, why can't it constrain the new credence about non-eternal matters? Some of the proposed norms are also incomplete: they are compatible with both steadfast halfing and thirding; the former is then ruled out on intuitive grounds. I've argued that the correct completion of the *ideal* norms would indeed lead to steadfast halfing.

Another apparently diachronic argument for thirding is that halfing would make Beauty vulnerable to a diachronic Dutch book – at least if she obeys Causal Decision Theory. (See [Arntzenius 2002], [Hitchcock 2004], [Draper and Pust 2008], and [Briggs 2010].) The arguments have been contested (see, e.g., [Bradley and Leitgeb 2006]), but I'm happy to concede that they work, and that they reveal a deviation from ideal diachronic rationality. Ideally, Beauty would obey shifted conditionalization throughout the course of the experiment. She would then not be vulnerable to any of the Dutch books that have been proposed. Her predicament requires her to deviate from the ideal. It's not a surprise that the deviation can be exploited.

What would be surprising is if thirding made Beauty invulnerable to diachronic Dutch books. But this has never been shown, and it isn't true. There is, in fact, a trivial Dutch book argument against thirding. If Beauty obeys thirding and bets in accordance with her credences, she would accept a deal on Sunday that pays \$8 if the coin lands heads and \$-7 if it lands tails. On Monday, she would accept

34 For parallel reasons, the expected accuracy standard does not support halfing in [Bostrom 2007]'s *Beauty the High Roller* or [Button et al. 2024]'s *Informed SB*, where halfing would be problematic. (It does support halfing in [Titelbaum 2008]'s *Technicolor Beauty*, answering the instability challenge in [Briggs 2010].)

another deal that pays $-9 in case of heads and $6 in case of tails. She would make a guaranteed loss of $1.

Of course, if Beauty knew that she is about to be exploited in this manner then the exploitation would not work, because the Monday offer would give Beauty information (that it is Monday) that would change her beliefs. I'm not sure if this is a legitimate complaint.[35] But it doesn't matter. We can tweak the setup to work around the information provided by the offers.[36] Let's assume that the second deal is offered on Monday and on Tuesday, but comes into effect only if Beauty accepts it on both occasions. Concretely, suppose after any Monday or Tuesday awakening, Beauty has the option of saying 'deal'. If the coin lands heads and she says 'deal' on Monday then $9 are deducted from her bank account on Wednesday. If the coin lands tails and she says 'deal' on Monday *and* on Tuesday, then $6 are deposited into her account on Wednesday (but only once). In all other cases, her bank account is left untouched. Now assume Beauty is a thirder, she knows all this, and her utility is measured by the balance in her bank account on Wednesday. According to Causal Decision Theory, whether Beauty should say 'deal' depends on her degree of belief, conditional on *Tails*, that she says 'deal' at the other awakening. If this degree of belief is at least ¾, saying 'deal' maximizes (causal) expected utility. I see no reason why she couldn't be confident (conditional on *Tails*) that she says 'deal' at the other awakening.[37] If she is, she ought to say 'deal', for a sure loss of $1.[38]

In sum, the best arguments for thirding rely on the evidentialist assumption that Beauty's Monday morning credence (at least with respect to self-locating matters) is fully determined by her present evidence. The assumption is popular, but it is not common ground. I, for one, reject it. If instead we ask how Beauty should update her previous credence, taking into account the possible information loss, Lewisian halfing emerges as the most plausible answer.

7 Conclusion: Evidentialism, ideality, and higher-order evidence

It has often been noted that classical conditionalization does not yield sensible results in cases that involve memory loss or self-locating information. Some have inferred that we should reject not just classical conditionalization, but all diachronic norms of rationality, falling back on purely synchronic norms like the evidentialist requirement to proportion one's beliefs to the present evidence. (See [Moss 2015], [Hedden 2015], [Meacham 2016], among many others.) In my view, this is an overreaction.

35 If it is, there can be no Dutch book argument for probabilistic coherence, since an incoherent agent may well become coherent (in the relevant respects) when faced with the chosen bets. Imagine, for example, an agent who gives credence ¼ both to the hypothesis H that they will be offered some bets today and also to its negation $\neg H$. How would you Dutch book them, if their credence in H and $\neg H$ would change to 1 and 0, respectively, as soon as they are offered any bets?

36 The tweaked setup goes back, in essence, to [Halpern 2006].

37 The causal decision problem has two equilibria, in the sense of [Skyrms 1990]: one in which Beauty is certain that she accepts the deal and one in which she is certain that she rejects it. In the "accept" equilibrium, the expected utility of Beauty's choice is $1. In the "reject" equilibrium, it is $0. Some causal decision theorists, such as [Weirich 1986] and [Arntzenius 2008], hold that only the better equilibrium is rationally permitted. This would mean that Beauty unequivocally has to say 'deal'. Other causal decision theorists, such as [Joyce 2012] and [Armendt 2019], allow for both choices.

38 Beauty also makes a sure loss in this setup if she follows Evidential Decision Theory.

I want to separate the two issues. To deal with self-locating information, we can adjust the rule of classical conditionalization. A promising alternative is shifted conditionalization. Both rules require perfect memory. So we still face the second issue: how should an agent update their beliefs if they are prone to losing information?

The answer is that it depends. The agent should try to control the damage, but what this means varies from case to case. Expected accuracy considerations provide one way to measure the damage. Even if the ideal update is not available, we can assess the available updates by their expected (future) accuracy.

When dealing with real agents, it is often difficult to establish whether an apparently irrational behaviour is an optimal adaptation to contextual constraints. I suspect that my own forgetfulness can't be justified as making optimal use of my cognitive resources. In the Sleeping Beauty problem, things are simpler, as we know exactly what constraints Beauty is under. I've argued that the optimal update that satisfies these constraints leads to Lewisian halfing. This is how Beauty *ought* to update her beliefs, in the "non-ideal" sense that takes into account the constraints of her situation.[39]

I have used the Sleeping Beauty problem as a test case because it brings out some further issues that are worth discussing. For example, it illustrates that an optimal belief update can lead to beliefs that are not proportioned to the later evidence. When Beauty wakes up on Monday, her evidence supports *Heads* to degree 1/3, yet the optimal update would leave her with credence 1/2. On reflection, it should be unsurprising that this kind of mismatch can occur. Diachronic norms constrain an agent's beliefs by their present evidence *and their earlier beliefs*. We should expect there to be cases where such norms clash with any principle that attempts to constrain an agent's beliefs by their present evidence alone.

Another important point *Sleeping Beauty* brings out is that fallback norms are needed not only if compliance with the ideal norms is impossible. Beauty may well be able to follow the ideal of shifted conditionalization throughout the experiment. It depends on whether the coin lands heads. Even if it lands tails, she is able to follow shifted conditionalization in the transition from Sunday to Monday. (Doing so would lead to steadfast halfing.) I have argued that Beauty should deviate from the ideal norm on Sunday night merely because she is unsure whether she will be able to follow it on Monday night. The reason why the mere possibility of cognitive impairment makes a difference is the limited state-dependence of epistemic norms. Beauty is subject to the same norms whether the coin lands heads or tails, although she only suffers from cognitive impairment in case of tails.

In light of this, it may be misleading to think of the standard Bayesian norms as norms of ideal rationality. Conditionalization and its shifted counterpart, for example, characterize the optimal update protocol for agents who happen to know that they are able to consistently follow this very protocol. It is odd to think that ideal rationality requires contingent knowledge of one's (present and future) cognitive abilities. Agents who lack such knowledge may still be ideally rational in any reasonable sense. That's why I've sometimes talked about compensatory norms as norms for non-ideal *circumstances*.

This brings me to a final lesson I want to draw. There has been a debate in recent years about how one should respond to evidence that one's cognitive capacities might be impaired. (See [Horowitz

39 I don't believe that 'ought' is genuinely ambiguous between and "ideal" and a "non-ideal" sense. Rather, the meaning of 'ought' seems have an argument for circumstances that are held fixed. See, for example, the classical treatment of [Kratzer 1977].

2022] for an overview.) When Beauty wakes up on Monday, she finds herself in this kind of quandary. Her evidence suggests that it may be Tuesday, in which case she has failed to properly retain and update yesterday's beliefs.

In the standard version of the story, Beauty already knows on Sunday about the experiment she is about to undergo. But this is immaterial. We could have reached the same conclusions if we had assumed that Beauty only learns about her predicament after awakening on Monday.[40] In this version of the story, Beauty initially wakes up confident that it is Monday and that she hasn't been awake since Sunday. When she learns about the experiment, she learns that her diachronic cognitive capacities may be impaired. Given that her evidence confers positive probability to $\mathit{Tails} \wedge \mathit{Tuesday}$, it also suggests that the formation of her present beliefs may have been affected by cognitive malfunction.

But suppose the coin has landed heads. Then the evidence is misleading. Steadfast halfing says that Beauty's "first-order" beliefs should be unaffected by her "higher-order" evidence: she should be certain that it is Monday and that nobody has tinkered with her memory.

I have argued that this is wrong. It is wrong because it would have led to an epistemically bad outcome if the coin had landed tails, and because the norms of rationality do not depend on the outcome of the coin toss. Beauty's misleading higher-order evidence makes a difference to her first-order beliefs.

Most of the debate over "higher-order evidence" has focussed on cases in which the relevant cognitive capacity is one of a priori reasoning. We are invited to imagine an agent who, for example, comes to believe a complex logical truth T on the basis of a proof, and then receives misleading evidence that their capacity of finding and assessing proofs is unreliable. We might expect that here, too, the higher-order evidence should affect the agent's first-order attitude towards T. But it's much harder to develop a clear model of this kind of situation.

The ideal agents of classical Bayesianism never come to believe logical truths on the basis of a proof. Ideal Bayesians have no use for a priori reasoning. They are certain of all logical truths without having gone through any proofs. If we want to assess how non-ideal agents should respond to evidence of their a priori fallibility, we first need to understand how positive credence can be assigned to logically impossible propositions, and how an agent with such incoherent credences should respond to evidence. In classical Bayesianism, evidence is closed under logical equivalence. If an agent receives evidence E, they also receive evidence $E \wedge T$, where T is any logical truth. To deal with our imperfection of seeing through the consequences of what we learn, we need a different conception of evidence and evidential support. All this raises hard questions, both technical and philosophical.

In the Sleeping Beauty problem, such problems don't arise. Beauty's predicament poses no threat to her probabilistic coherence. The only threat to her ideal rationality is the possible reset of her belief state on Monday night. We can evaluate her update dispositions without considering impossible worlds and non-monotonic concepts of evidence. In this respect, at least, the Sleeping Beauty problem is easy.

40 To model this cleanly, we may assume that Beauty gave negligible credence ϵ to the setup (S) on Sunday. The optimal update would then let her wake up with credence $\epsilon/2$ in $S \wedge \mathit{Heads}$ and $3\epsilon/4$ in $S \wedge \mathit{Monday}$. Conditioning on S yields Lewisian halfing.

References

Brad Armendt [2019]: "Causal Decision Theory and Decision Instability". *The Journal of Philosophy*, 116(5): 263–277

Frank Arntzenius [2002]: "Reflections on Sleeping Beauty". *Analysis*, 62(1): 53–62

— [2003]: "Some Problems for Conditionalization and Reflection". *Journal of Philosophy*, 100: 356–370

— [2008]: "No Regrets, or: Edith Piaf Revamps Decision Theory". *Erkenntnis*, 68: 277–297

Frank Arntzenius and Cian Dorr [2017]: "Self-Locating Priors and Cosmological Measures". In Khalil Chamcham, John Barrow, Simon Saunders and Joe Silk (Eds.) *The Philosophy of Cosmology,* Cambridge: Cambridge University Press

Nick Bostrom [2007]: "Sleeping Beauty and Self-Location: A Hybrid Model". *Synthese*, 157: 59–78

Darren Bradley [2011]: "Self-Location Is No Problem for Conditionalization". *Synthese*, 182: 393–411

Darren Bradley and Hannes Leitgeb [2006]: "When Betting Odds and Credences Come Apart: More Worries for Dutch Book Arguments". *Analysis*: 119–127

Rachael Briggs [2010]: "Putting a Value on Beauty". In T. Szabo Gendler and J. Hawthorne (Eds.) *Oxford Studies in Epistemology,* vol Vol. 3. Oxford: Oxford University Press

David Builes [2020]: "Time-Slice Rationality and Self-Locating Belief". *Philosophical Studies*, 177(10): 3033–3049

Tim Button, Daniel Rothschild and Levi Spectre [2024]: "Too Clever by Halving"

David Christensen [2007]: "Does Murphy's Law Apply in Epistemology? Self-doubt and Rational Ideals". *Oxford studies in epistemology*: 1

Joshua DiPaolo [2019]: "Second Best Epistemology: Fallibility and Normativity". *Philosophical Studies*, 176(8): 2043–2066

Cian Dorr [2002]: "Sleeping Beauty: In Defence of Elga". *Analysis*, 62: 292–296

Kai Draper and Joel Pust [2008]: "Diachronic Dutch Books and Sleeping Beauty". *Synthese*, 164(2): 281–287

John Earman [1992]: *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge (MA): MIT Press

Kenny Easwaran [2013]: "Expected Accuracy Supports Conditionalization—and Conglomerability and Reflection". *Philosophy of Science*, 80(1): 119–142

Adam Elga [2000]: “Self-Locating Belief and the Sleeping Beauty Problem”. *Analysis*, 60: 143–147

J. Dmitri Gallow [2021]: “Updating for Externalists”. *Noûs*, 55(3): 487–516

Hilary Greaves [2013]: “Epistemic Decision Theory”. *Mind*, 122(488): 915–952

Hilary Greaves and David Wallace [2006]: “Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility”. *Mind*, 115: 607–632

Ian Hacking [1967]: “Slightly More Realistic Personal Probability”. *Philosophy of Science*: 311–325

Joseph Halpern [2006]: “Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems”. In Tamar Gendler and John Hawthorne (Eds.) *Oxford Studies in Epistemology, Vol.1,* Oxford University Press, 111–142

Patrick Hawley [2013]: “Inertia, Optimism and Beauty”. *Noûs*, 47(1): 85–103

Brian Hedden [2015]: “Time-Slice Rationality”. *Mind*, 124(494): 449–491

Matthias Hild [1998]: “The Coherence Argument Against Conditionalization”. *Synthese*, 115(2): 229

Christopher Hitchcock [2004]: “Beauty and the Bets”. *Synthese*, 139: 405–420

Terry Horgan [2004]: “Sleeping Beauty Awakened: New Odds at the Dawn of the New Day”. *Analysis*, 64: 10–21

— [2008]: “Synchronic Bayesian Updating and the Sleeping Beauty Problem: Reply to Pust”. *Synthese*, 160: 155–159

Sophie Horowitz [2022]: “Higher-Order Evidence”. In Edward N. Zalta and Uri Nodelman (Eds.) *The Stanford Encyclopedia of Philosophy,* Metaphysics Research Lab, Stanford University, fall 2022 edition

Simon M. Huttegger [2013]: “In Defense of Reflection”. *Philosophy of Science*, 80(3): 413–433

Yoaav Isaacs, John Hawthorne and Jeffrey Sanford Russell [2022]: “Multiple Universes and Self-Locating Evidence”. *Philosophical Review*, 131(3): 241–294

Yoaav Isaacs and Jeffrey Sanford Russell [2023]: “Updating without Evidence”. *Noûs*, 57(3): 576–599

James Joyce [2009]: “Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief”. In F. Huber and C. Schmidt-Petri (Eds.) *Degrees of Belief,* Berlin: Springer, 263–300

— [2012]: “Regret and Instability in Causal Decision Theory”. *Synthese*, 187(1): 123–145

Brian Kierland and Bradley Monton [2005]: "Minimizing Inaccuracy for Self-Locating Beliefs". *Philosophy and Phenomenological Research*, 70(2): 384–395

Namjoong Kim [2009]: "Sleeping Beauty and Shifted Jeffrey Conditionalization". *Synthese*, 168(2): 295–312

Jason Konek and Benjamin A Levinstein [2019]: "The Foundations of Epistemic Decision Theory". *Mind*, 128(509): 69–107

Angelika Kratzer [1977]: "What 'must' and 'can' Must and Can Mean". *Linguistics and Philosophy*, 1(3): 337–355

Benjamin Anders Levinstein [2017]: "A Pragmatist's Guide to Epistemic Utility". *Philosophy of Science*, 84(4): 613–638

David Lewis [1979]: "Attitudes *De Dicto* and *De Se*". *The Philosophical Review*, 88: 513–543

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

Peter Lewis [2007]: "Quantum Sleeping Beauty". *Analysis*, 67: 59–65

Richard G. Lipsey and Kelvin Lancaster [1956]: "The General Theory of Second Best". *The review of economic studies*, 24(1): 11–32

Christopher Meacham [2008]: "Sleeping Beauty and the Dynamics of de Se Beliefs". *Philosophical Studies*: 245–269

— [2016]: "Ur-Priors, Conditionalization, and Ur-Prior Conditionalization". *Ergo*, 3: 444–402

Sarah Moss [2012]: "Updating as Communication". *Philosophy and Phenomenological Research*, 85(2): 225–248

— [2015]: "Time-Slice Epistemology and Action under Indeterminacy". *Oxford studies in epistemology*, 5: 172–94

Graham Oddie [1994]: "Harmony, Purity, Truth". *Mind*, 103(412): 451–472

— [1997]: "Conditionalization, Cogency, and Cognitive Value". *British Journal for the Philosophy of Science*, 48: 533–541

Richard Pettigrew [2016]: *Accuracy and the Laws of Credence*. Oxford University Press

Michele Piccione and Ariel Rubinstein [1997]: "On the Interpretation of Decision Problems with Imperfect Recall". *Games and Economic Behavior*, 20: 3–24

Joel Pust [2008]: "Horgan on Sleeping Beauty". *Synthese*, 160: 97–101

— [2023]: "No Double-Halfer Embarrassment: A Reply to Titelbaum". *Analytic Philosophy*, Forthcoming

Jeffrey Sanford Russell [2023]: "Sleeping Beauty's Evidence". In Maria Lasonen-Aarnio and Clayton M. Littlejohn (Eds.) *The Routledge Handbook of the Philosophy of Evidence,* Routledge

Miriam Schoenfield [2018]: "An Accuracy Based Approach to Higher Order Evidence". *Philosophy and Phenomenological Research*, 96(3): 690–715

Moritz Schulz [2010]: "The Dynamics of Indexical Belief". *Erkenntnis*, 72(3): 337–351

Wolfgang Schwarz [2015]: "Belief Update across Fission". *The British Journal for the Philosophy of Science*, 66(3): 659–682

— [2017]: "Diachronic Norms for Self-Locating Beliefs". *Ergo: An Open Access Journal of Philosophy*, 4: 709–738

Mattias Skipper and Jens Christian Bjerring [2022]: "Bayesianism for Non-ideal Agents". *Erkenntnis*, 87(1): 93–115

Brian Skyrms [1990]: *The Dynamics of Rational Deliberation*. Cambridge (Mass.): Harvard University Press

Wolfgang Spohn [2017]: "The Epistemology and Auto-Epistemology of Temporal Self-Location and Forgetfulness". *Ergo*, 4(13): 359–418

Julia Staffel [2019]: *Unsettled Thoughts: A Theory of Degrees of Rationality*. Oxford University Press

W. J. Talbott [1991]: "Two Principles of Bayesian Epistemology". *Philosophical Studies*, 62(2): 135–150

Michael G. Titelbaum [2008]: "The Relevance of Self-Locating Beliefs". *Philosophical Review*, 117(4): 555–606

— [2013]: *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford: Oxford University Press

Peter M. Todd and Gerd Ed Gigerenzer [2012]: *Ecological Rationality: Intelligence in the World.*. Oxford University Press

Bas C. van Fraassen [1984]: "Belief and the Will". *Journal of Philosophy*, 81(5): 235–256

Brian Weatherson [2015]: "Memory, Belief and Time". *Canadian Journal of Philosophy*, 45(5-6): 692–715

Paul Weirich [1986]: "Decisions in Dynamic Settings". In A. Fine and P. Machamer (Eds.) *PSA 1806,* East Lansing: Philosophy of Science Association, 438–449

Timothy Williamson [2000]: *Knowledge and Its Limits*. Oxford: Oxford University Press

Lyle Zynda [1996]: "Coherence as an Ideal of Rationality". *Synthese*, 109(2): 175–216