Changing minds in a changing world

Wolfgang Schwarz*

I defend a general rule for updating beliefs that takes into account both the impact of new evidence and changes in the subject's location. The rule combines standard conditioning with a "shifting" operation that moves the center of each doxastic possibility forward to the next point where information arrives. I show that well-known arguments for conditioning lead to this combination when centered information is taken into account. I also discuss how my proposal relates to other recent proposals, what results it delivers for puzzles like the Sleeping Beauty problem, and whether there are diachronic constraints on rational belief at all.

1 Introduction

As we make our way through the universe – by walking around town, by orbiting the sun, or simply by moving forward in time – we have to update our beliefs to keep track of our changing location. Right now I believe that it is Monday; sometime tonight, this attitude will fade and I will start believing that it is Tuesday. Philosophers disagree on how to model this kind of belief change. On the most straightforward account, we possess genuine information not only about the universe as a whole, but also about where and when in the universe we are. When the church bell strikes midnight, I can rule out alternative ways things might have been – not for the universe, but for *me, then*: it might have been earlier, or later.

* For helpful comments on earlier versions I would like to thank Jens Christian Bjerring, David Chalmers, John Cusbert, Alan Hájek, Namjoong Kim, Stephan Leuenberger, Weng Hong Tang, Michael Titelbaum, J. Robert G. Williams, two anonymous referees, and the audience of a PhilSoc seminar at the Australian National University in February 2008. Several points made in this paper, especially in section 6, are also made in [Meacham 2010], which appeared while this paper was under review.

A maximally specific way things might have been for an individual at a time is a *centered (possible) world*. A less specific way – a *centered proposition* – can be modelled as a class of centered worlds. As [Lewis 1979] points out, we don't need a special treatment for uncentered worlds and propositions, since every way a universe might be determines a way things might be for an individual at a time: to be such that the universe is so-and-so. From now on, when I use 'world' and 'proposition', I always mean centered worlds and centered propositions.

Since propositions (so understood) can change their truth value over time, it is possible to believe a proposition A at one time and believe not-A at a later time, and still think that the previous belief was true. Today's belief that it is Monday is not in tension with tomorrow's belief that it is Tuesday. By contrast, when we revise our beliefs in response to new information, we typically consider our previous beliefs to be false, or less accurate. These two kinds of belief change have been extensively studied in stochastic control theory and the AGM school of formal epistemology (see e.g. [Kumar and Varaiya 1986], [Katsuno and Mendelzon 1991]); outside these areas however, the update process characteristic of "self-locating" beliefs is still largely ignored. Its relevance has only recently surfaced in the wake of the Sleeping Beauty problem.

I will defend an update rule that incorporates both kinds of belief change, loosely building on ideas from control theory. The rule combines conditioning with a "shifting" operation that moves the center of each doxastic possibility forward to the next point where information arrives. Section 2 introduces and motivates the basic proposal; section 3 looks at some consequences for an agent's attitude towards her past and future beliefs. In section 4, I show that several traditional arguments for conditioning support the combination with shifting once centered information is taken into account. Section 5 compares my proposal to a different approach that has become popular in the Sleeping Beauty debate. In section 6, I discuss a tension that arises between diachronic constraints on rationality and the idea that rational belief cannot be constrained by matters inaccessible to the subject.

Although I work in the centered worlds framework, my proposal should also be useful in alternative frameworks. For instance, [Perry 1979] and others have argued that belief should be understood as a three-place relation between a subject, an uncentered content, and a *mode of presentation* which encodes information about the subject's location. To make sense of uncertainty and

evidence about one's location, degrees of belief should then be assigned not only to contents, but also to modes. Given the familiar representation of modes as functions from centered worlds to uncentered contents, a probability distribution over modes (or mode-content pairs) determines a probability distribution over centered worlds by diagonalisation. My proposal would then concern the dynamics of these diagonal probabilities.

2 Shifted conditioning

In the centered worlds framework, a belief state is modelled by a probability distribution over centered worlds. When new evidence comes in, this distribution gets revised. The classical rule for such revisions, *conditioning*, presupposes that the new evidence makes the subject certain of some evidence proposition E; any other proposition A then gets its probability adjusted to its old probability conditional on E:

(C) $\quad P_2(A) = P_1(A \,|\, E).$

To see why this does not adequately handle centered information, consider a simple example.

> **The Litmus Test.** You dip a piece of white litmus paper into a beaker which you suspect to contain acid. The paper turns red.

Let P_1 be your credence function just before you see the paper turn red, and let E be the information you then receive. P_1 assigns high probability to worlds where the paper is about to turn red and somewhat lower probability to worlds where it is about to turn blue. Very low probability goes to worlds where the paper has already turned red. Perhaps you cannot completely rule out such worlds: you might be hallucinating that the paper is white. Perhaps you can't even rule out worlds where the paper *looks* red, so that you are somehow mistaken even about the paper's appearance. But if this is how things are, then very strange things must be going on, involving evil demons or other malignant forces.

A moment later, you observe that the paper is red. Conditioning would have you move all your credence to worlds where the paper looks red, in proportion to their previous probability. You would become convinced that very strange

things are going on. Your actual response, of course, is to believe that it is now somewhat later than before and that the paper has in the meantime turned red. Conditioning is too conservative with your self-locating beliefs. Since you expected the paper to turn red, you should not hold on to your belief that either the paper is white or you are being fooled by evil demons. (Notice that this problem does not exploit the assumption that agents are absolutely certain about their present evidence.)

Conditioning is a reasonable strategy for revising beliefs about the world in the light of new information *about that same world*. But as we move through space and time, we leave our old (centered) worlds behind and enter new ones – new worlds where what was true before may now be false. We need an update policy that can take such changes into account.

Before I present my proposal, let me explain what I mean by an *update policy*. Formally, an update policy is a mapping from a prior probability function P_1 and an evidence proposition E to a posterior probability function P_2. An agent *follows* an update policy (at a given time) if her credence equals the result of that function applied to her previous credence and her total new evidence. A more general notion is required if the evidence does not confer certainty on any proposition; I will return to such generalisations at the end of this section.

What is the agent's 'previous credence function'? It can't be an arbitrary credence from any earlier time. A policy that non-trivially operates on P_1 will likely yield different outputs when supplied with different inputs. Since there can be only one new credence function P_2, we should not allow P_1 to come from arbitrary earlier times. P_1 should be the credence function from just before the new evidence arrived. ("Just before" relative to the agent's personal time; for a time traveller, the "just before" state may lie in the distant past or future.) When we discuss particular examples, we may of course ignore times at which no information relevant to the propositions of interest arrives, just as we commonly ignore irrelevant aspects of the total new evidence.

One could perhaps use arbitrary earlier credence if evidence was cumulative: if later evidence always contained full information about all previous evidence, including the order in which it arrived. But this is a rather unrealistic assumption; I want to allow for agents who are not so fortunate as to always have complete evidence about everything they have ever learnt.

What if there is a *continuous* stream of (relevant) evidence? That is, what if

for any previous time there is an even closer time at which relevant information arrives? If evidence does not accumulate over these intervals, we will miss information, no matter which credence we take as P_1. A discrete update model can then only approximate the optimal update process. Again, for practical applications, the approximation will often be good enough. But if we are interested in the optimal update itself, we have to understand update policies more generally as operations that map an old probability, a *time interval* and a *stream of evidence* to a new probability. The policy I will defend can easily be generalised in this way (along the lines of [LaValle 2006: 589–598]). For the sake of simplicity, I will here stick to discrete updates.

So assume that for any centered world with positive credence, there is a unique "next" point where (relevant) information arrives. Let '$\succ A$' express the proposition that A will be true at the next point where evidence arrives: $\succ A$ is true at world w iff A is true at the next point at w where information comes in. On the policy I recommend, the new credence in A after learning E then equals the previous conditional credence in $\succ A$ conditioned on $\succ E$:

(SC) $\quad P_2(A) = P_1(\succ A \,|\, \succ E)$.

The shifting operator $\succ$ induces a transformation on the space of probability functions, mapping the credence P_1 to the *shifted credence* $P_1{}^{\succ}$, with $P_1{}^{\succ}(A) = P_1(\succ A)$. Shifting and conditioning commute: if you first condition P_1 on $\succ E$ and then shift, the result is the same as if you first shift and then condition on E. Hence the new credence also equals the shifted previous credence conditioned on the new evidence:

(SC) $\quad P_2(A) = P_1{}^{\succ}(A \,|\, E)$.

If A is certain not to change its truth-value in the foreseeable future, then $P_1(\succ A) = P_1(A)$. If both A and E have this property, then (SC) reduces to (C). In this sense, (SC) is a generalisation of conditioning, adjusted to handle transient propositions.

For a simple example, suppose at t_1 you believe that the sun is shining, and your evidence at t_2 is neutral on this matter. Then at t_2 you will believe that the sun was shining just before the present evidence arrived. You do not hold fixed your self-locating beliefs; you don't assume that things are still exactly the way they were before. Nor do you completely ignore those beliefs. You

assume – tentatively, and subject to revision in the light of new evidence – that the previous beliefs represent how things were a moment earlier.

Consider again *The Litmus Test*. Let t_1 be the time when you start dipping the paper. Suppose at this point, 60% of your credence P_1 goes to worlds where the liquid is an acid and 40% to other possibilities. Let's say your credence that the paper is about to turn red is 0.9 conditional on the first kind of situation, and 0.1 conditional on the second. By Bayes' Theorem, your credence $P_1(A \mid \succ R)$ that the liquid is an acid given that it is about to turn red is

$$\begin{aligned} P_1(A \mid \succ R) &= \frac{P_1(\succ R \mid A) \times P_1(A)}{P_1(\succ R \mid A) \times P_1(A) + P_1(\succ R \mid \neg A) \times P_1(\neg A)} \\ &= \frac{0.9 \times 0.6}{0.9 \times 0.6 + 0.1 \times 0.4} \\ &\approx 0.93. \end{aligned}$$

Since you don't assume that the liquid can change its acidity, this is also the value of $P_1(\succ A \mid \succ R)$. So if at t_2 your relevant new evidence is R, then by (SC), your new credence in A is $P_2(A) = P_1(\succ A \mid \succ R) \approx 0.93$. You end up 93% confident that the liquid is an acid. No probability is moved to evil demon worlds.

It may help to picture this process in a table.

		$A\&R$	$\neg A\&R$	$A\&\neg R$	$\neg A\&\neg R$
0.6	$A\&\neg R$	0.9	0	0.1	0
0.4	$\neg A\&\neg R$	0	0.1	0	0.9
Shifting:		*0.54*	*0.04*	*0.06*	*0.36*
Conditioning:		*0.93*	*0.07*	*0*	*0*

The italicised values at the left are the old probabilities, the values at the bottom the new ones. The body of the table (above the line) contains the *transition probabilities*: the fraction of the old credence in the proposition marking the row that gets transferred to the proposition marking the column. For example, given that the liquid is an acid (we're in the top row), the probability that the present state will turn into one where the paper is red is 0.9. The values in the 'Shifting' row are calculated by adding up the numbers in each column multiplied by the old probability of the relevant row. Under 'Conditioning', the possibilities that are ruled out by the evidence are set to zero and the other probabilities renormalised.

Transition probabilities are a familiar parameter in stochastic control theory, but they play a slightly different role there (see e.g. [Kumar and Varaiya 1986], [LaValle 2006: part III]). In control theory, subjective probability is defined not over centered worlds, but over fragments of stages of worlds, called *states*. States do not contain information about the past or the future. To evaluate their options, agents therefore need not only a probability distribution over states, but also an idea of how these states will evolve. This is represented by the transition probabilities. In the present framework, transition probabilities are derived values, determined by the ordinary subjective probabilities. If state s_0 could develop into s_1 or s_2, then we have two kinds of worlds all along, one developing into s_1 and one into s_2. The transition probability between A and B is simply the agent's credence that B will be true at the next point when information comes in given that A is true now.

Since shifted conditioning combines shifting with conditioning, it can easily be adjusted to other revision rules. For example, suppose the new evidence determines a distribution of probabilities $x_1, \ldots, x_n$ over a partition of propositions $E_1, \ldots, E_n$, rather than making any particular proposition certain. According to *Jeffrey conditioning* (see [Jeffrey 1983: 164–169]), the new probability P_2 is then given by

$$\text{(JC)} \quad P_2(A) = \sum_i P_1(A \,|\, E_i) \times x_i.$$

Replacing the conditioning step in (SC) with Jeffrey conditioning, we get

$$\text{(SJC)} \quad P_2(A) = \sum_i P_1(\succ A \,|\, \succ E_i) \times x_i.$$

Similar adjustments are possible for other variations of conditioning.

If we skip the conditioning step and apply shifting directly to the shifted probabilities, we get the double-shifted probability ${P_1}^{2\succ}$, the triple-shifted probability ${P_1}^{3\succ}$, etc. Intuitively, ${P_1}^{n\succ}(A)$ is the t_1 credence that A will be the case after n intakes of new information. I will write '$\succ_n A$' for this proposition (that A will be the case after n intakes of new information). So ${P_1}^{n\succ}(A) = P_1(\succ_n A)$. As we will see next, $P_1(\succ_n A)$ is also the agent's expectation of her future credence in A.

3 Shifted Reflection

Call an agent *self-aware* if she knows with certainty what update policy she follows and what her present credence is. Self-aware agents who follow conditioning have the characteristic property that their current credence matches the expectation of their future credence (see [Goldstein 1983], [van Fraassen 1984]). This property is known as *Reflection*:

$$\text{(R)} \qquad P_1(A) = \sum_x P_1(P_2(A) = x) \times x.$$

Informally, to satisfy Reflection means to trust one's (expected) future judgement. This kind of trust is evidently absurd if propositions can change their truth-value: tomorrow I will probably believe that it is Tuesday, but this does not mean that I should already believe now that it is Tuesday. If centered propositions are in play, we have to distinguish between assuming that a belief with content A is true, and assuming A. Trusting someone who believes that it is Tuesday is to assume that their belief is true; but this is not the same as assuming that it is Tuesday. If the trusted subject is known to be one day ahead, it rather means assuming that it is Tuesday *tomorrow*. In general, to trust the judgement of your successor is to satisfy a principle of *Shifted Reflection*:

$$\text{(SR)} \quad P_1(\succ A) = \sum_x P_1(P_2(A) = x) \times x.$$

Shifted Reflection characterises self-aware agents who follow (SC).[1]

Like (R), (SR) contains 'P_2' in the scope of 'P_1'. Since this is an intensional context, it matters how P_2 is presented. Suppose you are uncertain whether it is Monday or Tuesday, and you know that on Wednesday morning you will next find out what day it is. In fact it is Tuesday; so we can refer to tomorrow's credence as 'your credence on Wednesday'. But on this way of presenting P_2, you do not satisfy (SR): your present credence in it being Wednesday *tomorrow* may be 1/2 even though your expected *Wednesday* credence in it being Wednesday

1 Proof: let $E_1, \ldots, E_n$ be a partition of the evidence you might receive at t_2 such that E and E' fall in the same cell of the partition iff $P_1(\succ A \,|\, \succ E) = P_1(\succ A \,|\, \succ E')$. For each $i \leq n$, let $x_i = P_1(\succ A \,|\, \succ E)$, where E is any member of E_i. By the law of total probability, $P_1(\succ A) = \sum_i P_1(\succ E_i) \times x_i$. If you know that you update by (SC), $P_1(\succ E_i \leftrightarrow P_2(A) = x_i) = 1$. Hence $P_1(\succ A) = \sum_i P_1(P_2(A) = x_i) \times x_i$, and you satisfy (SR).

is 1. In general, (SR) only holds if P_2 is presented as 'my credence at the next point when information comes in'. A more accurate formulation would therefore go like this: $P_1(\succ A) = \sum_x P_1(\succ P(A) = x) \times x$. More generally, using the n-shifted probability function from the previous section,

$$\text{(SR)} \quad P_1(\succ_n A) = \sum_x P_1(\succ_n P(A) = x) \times x.$$

A nice illustration of these issues, which may also help to further clarify the application of (SC), is Frank Arntzenius's [2003] story of the prisoner. I will look at the following variation.

> **The prisoner.** A prisoner is waiting in her cell while a jury decides whether she will be executed or banished. If she faces execution, the lights in her cell get switched off at midnight. Aware of this arrangement, the prisoner falls into a restless sleep from which she briefly awakens at several points throughout the night. At each awakening, she finds the lights in her cell still on.

To simplify the model, assume that the prisoner falls asleep at 8pm, and at this point still knows what time it is. She also knows that each sleep phase takes either one hour or two. Her initial credence is distributed at follows.

8pm	8pm	9pm	10pm	11pm	12am	1am	2am
Execution	1/2	0	0	0	0	0	0
Banishment	1/2	0	0	0	0	0	0

Since she does not know when she will wake up next, the two open possibilities divide into four sub-possibilities, depending on whether the next sleep phase takes one hour or two. After the first awakening, her credence in any of these four possibilities is shifted to the corresponding combination of *Execution*/*Banishment* with *9pm*/*10pm*.

1st awak.	8pm	9pm	10pm	11pm	12am	1am	2am
Execution	0	1/4	1/4	0	0	0	0
Banishment	0	1/4	1/4	0	0	0	0

At the second awakening, it could be 10pm, 11pm, or 12am. The combination *Execution & 12am* is excluded by the evidence that the lights are still on. By (SC), the new credence is[2]

2 Here is the update table:

2nd awak.	8pm	9pm	10pm	11pm	12am	1am	2am
Execution	0	0	1/7	2/7	0	0	0
Banishment	0	0	1/7	2/7	1/7	0	0

At the third awakening, the probability of *Execution* has further decreased:

3rd awak.	8pm	9pm	10pm	11pm	12am	1am	2am
Execution	0	0	0	1/9	0	0	0
Banishment	0	0	0	1/9	1/3	1/3	1/9

At the fourth awakening, she is certain that she will be banished.

The prisoner's credence gradually spreads over larger and larger intervals of time: it spans n hours after the nth awakening. This is not due to any cognitive failures, but simply to the fact that she lacks information about how much time passes between the awakenings. Her situation resembles that of a time traveller who enters a time machine not knowing how far it will take her into the past or the future.

As Arntzenius points out, the prisoner appears to violate Reflection. For suppose she is aware of her update policy; then she knows at 8pm that by 11pm, her credence in *Banishment* will be either 1/2 or 4/7 or 8/9, depending on whether there will be one, two or three awakenings until then. The expectation of her 11pm credence is therefore greater than 1/2. In general, whatever her credence in *Banishment* is at 8pm, the expectation of her 11pm credence is greater (unless the 8pm credences is 1). The prisoner cannot trust her future self!

The problem here is that the future credence is picked out in an illegitimate way, as the credence *at 11pm*. By contrast, consider the prisoner's expectations about her beliefs *after two awakenings*. Her credence in *Banishment* will then either be 4/7 (if the lights are still on) or 0 (if the lights are off). Since the

		E 10	E 11	E 12	B 10	B 11	B 12
1/4	E 9	1/2	1/2	0	0	0	0
1/4	E 10	0	1/2	1/2	0	0	0
1/4	B 9	0	0	0	1/2	1/2	0
1/4	B 10	0	0	0	0	1/2	1/2
Shifting:		*1/8*	*1/4*	*1/8*	*1/8*	*1/4*	*1/8*
Conditioning:		*1/7*	*2/7*	*0*	*1/7*	*2/7*	*1/7*

probability that the lights will be off by the second awakening is 1/8, the expectation of the future credence in *Banishment* is $4/7 \times 7/8 = 1/2$.[3]

To understand why Reflection fails if the future credence is picked out by a definite time, note that the prisoner knows something about her situation at 11pm that she will not know when she is there: that it is located at 11pm. If you trust someone's judgement while possessing information they lack, you should not endorse their unconditional judgement, but their (expected) *conditional* judgement, conditional on the further information you possess about their situation. The prisoner's 11pm credence in *Banishment* conditional on it being 11pm is 1/2. Curiously, the "further information" the prisoner has about her future self is not some interesting fact about the world. The prisoner knows that her 11pm successor is located at 11pm simply because that is how she picks her out.

Reflection relates the present credence to the expected future credence. We can also relate the present credence to earlier credence. In this case, it is obviously quite common that the agent has relevant information that she previously lacked. Among self-aware agents who follow conditioning, the later credence therefore matches the expectation of the previous credence conditional on the new evidence:

$$(\text{Я}) \qquad P_2(A) = \sum_x P_2(P_1(A \,|\, E) = x) \times x.$$

The shifted version is

$$(\text{SЯ}) \qquad P_2(A) = \sum_x P_2(P_1(\succ A \,|\, \succ E) = x) \times x.$$

As before, it is easy to verify that self-aware agents who follow (C) or (SC) satisfy the corresponding principle of inverse Reflection – though again we have

3 Arntzenius mentions a related puzzle. Suppose the prisoner knows in advance what her credence will be at 11pm. Then she could use her beliefs as a clock: she could figure out whether it is 11pm merely by introspecting if she has the relevant beliefs. Does this mean that self-awareness is incompatible with losing track of the time? No. If you know what update policy you follow and what evidence you will receive, then you know what your credence will be at each point when information arrives. But you need not know how these "points when information arrives" map onto what is measured by our clocks. (Again, think of the time traveler who does not know where the time machine will take her.)

to be careful that P_1 is picked out in the right way: as the credence just before the present information came in.[4]

4 Conditioning revisited

Many arguments have been put forward in support of conditioning, showing that it is the only update rule with certain desirable features. Once we allow propositions to change their truth-value, it turns out that these arguments actually support shifted conditioning. I will demonstrate this for three well-known, and hopefully representative examples: an argument from coherence, an argument from Reflection, and an argument from minimal revision.

The first argument, due to Lewis [1999] and first published in [Teller 1973], might be summarised as follows. Imagine you know that tomorrow you will find out (for certain) whether or not E obtains. In response to this, you plan to update your credence from P_1 to either P^E or $P^{\neg E}$ accordingly. Let A be any proposition, and consider an arrangement that will cost you a certain amount of money x tomorrow if it turns out that $E \,\&\, \neg A$, and that will pay you $1 - x$ if it turns out that $E \,\&\, A$. (You neither gain nor lose if $\neg E$.) For what values of x do you judge this arrangement to have positive expected payoff for your future self?

On the one hand, the expected payoff is $-x \times P_1(E \,\&\, \neg A) + (1-x) \times P_1(E \,\&\, A)$, which is greater than 0 iff $x < P_1(E \,\&\, A)/P_1(E)$. On the other hand, if tomorrow you find out E and thus update your credence to P^E, the arrangement will be worth $-x \times P^E(\neg A) + (1 - x) \times P^E(A)$ to you; so today's estimate of tomorrow's value is $P_1(E)$ times that amount. This is positive iff $x < P^E(A)$. The two answers are compatible only if $P^E(A)$ equals $P_1(E \,\&\, A)/P_1(E)$. Hence to avoid having "contradictory opinions about the expected value of the very same transaction" [Lewis 1999: 405], you should plan to update your credence by conditioning.

More precisely, if $P^E(A)$ comes apart from $P_1(A \,|\, E)$, you effectively plan to update your credence in such a way that your future self will be mistaken (by

4 Proof for (SC): suppose you have just received the information E. Your new credence y in A then equals your previous credence in $\succ A$ conditional on $\succ E$. Being aware of both your present credence and your update policy, you can conclude that your previous credence in $\succ A$ conditional on $\succ E$ was y. So $P_2(P_1(\succ A \mid \succ E) = y) = 1$ and $P_2(A) = \sum_x P_2(P_1(\succ A \,|\, \succ E) = x) \times x$.

your present lights) about the expected cost of the arrangement. *Pace* Lewis, this is not quite a contradictory state of mind, but it is certainly peculiar. As a corollary, you will be susceptible to a Dutch Book: a clever bookie who knows nothing more than you could make a sure profit by selling you bets on the expected cost today and tomorrow in combination with a low stake bet against E.

If propositions can change their truth-value, this argument becomes invalid: the arrangement's outcome depends on whether or not E and A are true *tomorrow*; but in calculating the expected payoff as $-x \times P_1(E \,\&\, \neg A) + (1 - x) \times P_1(E \,\&\, A)$, we have used the probability of E and A *now*. If these propositions can change their truth-value, we get the wrong result. For example, if tomorrow I find out that it is sunny, I will come to believe that the washing in the garden is dry. Nevertheless, since I only just hung it out, my current conditional credence in the washing being dry given that it is sunny is rather low. What is high is my conditional credence in the washing being dry *tomorrow* given that it is sunny *tomorrow*. The arrangement's actual expected payoff is $-x \times P_1(\succ (E \,\&\, \neg A)) + (1 - x) \times P_1(\succ (E \,\&\, A))$, which is greater than zero iff $x < P_1(\succ (E \,\&\, A))/P_1(\succ E)$. Thus what the coherence argument really shows is that your updated credence $P^E(A)$ should equal your conditional *shifted* credence $P_1(\succ A \,|\, \succ E)$.

The next argument is an argument from Reflection. This time, we start with the assumption that under ideal conditions, one should trust the judgement of one's better-informed future self, as expressed by the Reflection principle

$$\text{(R)} \qquad P_1(A) = \sum_x P_1(P_2(A) = x) \times x.$$

As [van Fraassen 1999] shows, if an agent satisfies (R), then (under further weak assumptions) they cannot plan to update their credence by any rule other than conditioning.

Van Fraassen's proof remains valid if centered propositions are allowed, but Reflection itself becomes implausible. As argued in the previous section, trusting one's future self is then better expressed as

$$\text{(SR)} \quad P_1(\succ A) = \sum_x P_1(\succ P(A) = x) \times x.$$

(SR) is supported by a Dutch Book argument very similar to Lewis's; but here we just take it as a starting point. Following van Fraassen, we can turn it into

another argument for (SC): suppose an agent satisfies (SR) and is about to learn (for certain) whether E or $\neg E$, upon which she will update her credence to P^E or $P^{\neg E}$ accordingly. Suppose also she knows this. Then for any world w, $P_1(\succ w) = P_1(\succ E) \times P^E(w) + P_1(\succ \neg E) \times P^{\neg E}(w)$. If E is true at w, then $P^{\neg E}(w) = 0$ and $P_1(\succ w) = P_1(\succ E) \times P^E(w)$. Moreover, $P_1(\succ w) = P_1(\succ E) \times P_1(\succ (w\&E))/P_1(\succ E)$; hence $P^E(w) = P_1(\succ (w\&E))/P_1(\succ E) = P_1(\succ w \,|\, \succ E)$. On the other hand, if E is not true at w, then $P^E(w) = 0$ and $P_1(\succ w \,|\, \succ E) = 0$, so again $P^E(w) = P_1(\succ w \,|\, \succ E)$. So P^E results from P_1 by (SC).

Finally, I want to look at an argument from minimal revision. Suppose you consider two theories A and B to be equally probable. If both theories predict E, then after finding out that E (and nothing else), you should not judge A to be more probable than B. That is,

(MR) if A and B entail E, and $P_1(A) = P_1(B)$, then $P^E(A) = P^E(B)$.

[Teller 1973: 225–232] proves that conditioning is the only general rule for P^E that satisfies this constraint. (See [Williams 1980] and [Diaconis and Zabell 1982] for related results.)

(MR) entails that in the absence of relevant evidence, credence does not change. Again, this is not quite right if the world itself can change. If A says that E is true now and false afterwards, while B says that E is always true, then A and B both entail E; but when in a few moments you find out E, you have found strong evidence against A, and not against B. We should therefore replace (MR) by a shifted version:

(SMR) If A and B entail E, and $P_1(\succ A) = P_1(\succ B)$, then $P^E(A) = P^E(B)$.

For uncentered A and E, (SMR) reduces to (MR), just as (SR) reduces to (R).

It is easy to show that shifted conditioning satisfies (SMR): if A and B entail E, then $P_1(\succ (A\&E)) = P_1(\succ A)$ and $P_1(\succ (B\&E)) = P_1(\succ B)$, hence if $P_1(\succ A) = P_1(\succ B)$, then $P_1(\succ (A\&E))/P_1(\succ E) = P_1(\succ (B\&E))/P_1(\succ E)$. The converse, that *no other* policy satisfies (SMR), can be proved by straightforward but tedious adaptation of the proof in [Teller 1973].

5 Indirect conditioning

I now want to compare shifted conditioning to a family of alternatives that have become popular in the debate on Sleeping Beauty. The basic idea is to use old-fashioned conditioning, but somehow restrict it to uncentered propositions. Return once more to *The Litmus Test*, and consider the proposition E that specifies your total evidence after you have dipped the paper into the liquid. E is a very rich proposition. It contains information not just about the paper's colour, but also about the beaker, the table, the lighting, your state of hunger, and so on. You may reasonably think that E is true at most once in the history of the universe: other people at other places or times may also dip papers into beakers, but unless their environment and inner state mysteriously duplicates your present situation, the total evidence they receive will be different. So suppose you rule out worlds where E is true more than once. Your uncentered beliefs are then divided between worlds where E occurs once and worlds where E occurs never. By learning E, you can rule out worlds of the first kind. But you also know where in the remaining worlds you are: you must be at the very place and time where E occurs.

To make this precise, let '*somewhere* A' be true at a world w iff A is true at some place and time in the universe of w. '*Somewhere* A' is the strongest uncentered proposition entailed by A – uncentered insofar as it does not distinguish between different centers within the same universe. Let 'A *at* B' be shorthand for '*somewhere* B, and *everywhere*, $B \supset A$', where '*everywhere*' abbreviates 'not *somewhere* not'. Thus 'A *at* B' is true iff A holds at every place in the world where B holds, and there is at least one such place. Figure 1 should help to clarify these notions. Assuming that P_1 gives zero credence to worlds where E is true more than once, we can then express the policy of *indirect conditioning* as follows:

(IC) $\quad P_2(A) = P_1(A \textit{ at } E \mid \textit{somewhere } E).$

This takes the initial credence function P_1, rules out all worlds where E occurs nowhere, and moves the center in the remaining worlds to the point where E occurs (of which, by assumption, there will never be more than one). Renormalising yields the new probability P_2. In other words, (IC) conditions the uncentered fragment of P_1 on the uncentered fragment of E and re-introduces the centers as the point where E is true. In very broad outline, this is the policy

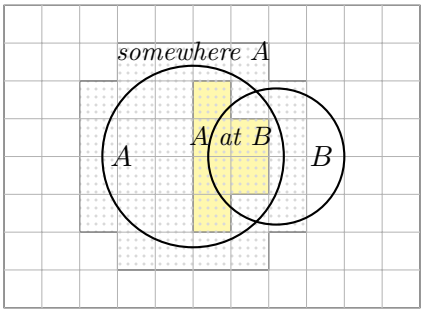

Figure 1: **uncentered propositions in centered logical space.** The grid partitions the space of worlds by the worldmate relation: two worlds are in the same cell iff they agree in what they say about the universe as a whole. '*Somewhere* A' and 'A *at* B' express uncentered propositions about the universe, even though A and B are centered.

recommended in [Halpern 2006], [Meacham 2008], [Titelbaum 2008], [Kim 2009], and, though less explicitly, [Elga 2000] and [Elga 2004].

Here is another way to motivate the proposal. Suppose at time t_2 you are confident that a certain self-locating proposition L is true here and now, and nowhere else. (Think of L as something like "it is 8 am on 12 April 2009 and I am standing at the corner of King's Court, Southwark, London".) Any proposition A can then be mapped onto the corresponding uncentered proposition 'A *at* L' without affecting its probability at t_2. If uncentered probabilities evolve by conditioning, this determines the new probability for every proposition:

(IC′) $\quad P_2(A) = P_2(A \text{ at } L) = P_1(A \text{ at } L \mid E \text{ at } L).$

To apply (IC′), we need a suitable self-locating certainty L. Presumably L must come from the new evidence; at least it can't have been learnt by (IC′). But if E entails L, and possibilities with multiple occurrences of E are ruled out, then 'E *at* L' reduces to '*somewhere* E'; so (IC′) is a special case of (IC). Moreover, if we are liberal about the propositions that can take the place of L, we can always plug in E itself, making (IC′) equivalent to (IC).

There is an obvious resemblance between (SC) and (IC′). Both apply conditioning to a shifted transformation of the previous credence. (IC′) shifts the

probability of any world w to wherever L holds within the same universe; (SC) shifts it to the next point at w where information comes in.[5]

To see how (SC) and (IC) can come apart, assume that A is known not to change its truth-value over time. According to shifted conditioning, an evidence proposition E raises the probability of A only if A makes it sufficiently likely that E will be encountered *next*: $P_2(A) > P_1(A)$ iff $P_1(A \mid \succ E) > P_1(A)$. According to indirect conditioning, E confirms A as long as A makes it likely that E occurs *at some point or other* in the history of the world: $P_2(A) > P_1(A)$ iff $P_1(A \mid \textit{somewhere } E) > P_1(A)$. For example, consider the following hypothesis.

> **Eternal Life.** You will live a very long life, in the course of which you will have almost every humanly possible experience.

Your current credence in *Eternal Life* is presumably low. However, let E be any experience you may have later today. Conditional on *Eternal Life*, the probability of E occurring at some point or other is very high, whereas it is much lower on more normal assumptions about your life. If you follow (IC), any experience whatsoever will increase your credence in *Eternal Life*.

(SC) does not have this unwelcome consequence, since it only considers what *Eternal Life* predicts about your very next experience, on which the hypothesis is largely silent.[6]

For another example, consider an Everettian interpretation of quantum mechanics on which the universe constantly branches, with every possible outcome of every chance process occurring in some branch of the universe. To avoid multiple occurrences of the evidence within a world, pretend that your evidence is specific enough to always tell you at which branch you're located. Now suppose you are watching the clouds, and cloud formation is a chancy process. On the Everettian theory, it was certain beforehand that in some branch of the universe the clouds would gather in exactly the way you find them. On alternative, "collapse" interpretations of quantum mechanics, the clouds could well have gathered only in different ways. Your evidence therefore

5 [Kim 2009] even calls his version of (IC′) "Shifted Strict Conditionalization". Understanding (SC) as an improved version of (IC′) may justify keeping the label.

6 By 'experience' I here mean something like sensory experience, not things like *being killed in World War I*. The example works best on a narrow conception of evidence, on which evidence is closely aligned with (sensory or memory) experience. Otherwise parts of your evidence that go beyond experience might e.g. tell against *Eternal Life*.

never rules out any previously possible Everett universe, while it constantly rules out previously possible universes with collapse. If you follow (IC), a few hours of observing the clouds should strongly increase your credence in the Everett interpretation.

In the previous section, we also saw that agents who deviate from (SC) must violate certain principles of rationality such as (shifted) Reflection, Minimal Revision, and immunity to Dutch Books. This is illustrated again by the two examples. If you follow (IC), you know that any observation you make today will increase your confidence in *Eternal Life*; hence your current credence in this proposition does not equal the expectation of your better-informed future credence: you violate Reflection. (*Eternal Life* does not change its truth-value over time, so the reasons for moving to shifted Reflection do not apply.) Moreover, if you bet in accordance with your beliefs, I can Dutch Book you by selling you a wager against *Eternal Life* now and buying it back at a reduced price later.

The root of these problems is that indirect conditioning is too revisionary about self-locating information: it tells your later self to dismiss your former beliefs about where in the world you are, even if the relevant self-locating propositions were certain not to change their truth-value. If before looking out of the window you knew that you live in the 21st century on Earth, (IC) completely ignores this knowledge and tries to establish your new location from scratch, based on the evidence you get from looking out of the window.

Yet more trouble emerges if we allow for worlds where the evidence occurs more than once. As presented so far, (IC) then collapses into contradiction. For suppose at t_1 you believe that E will occur on Monday and then again on Tuesday. By (IC), learning E should make you certain that it is not Monday, as the P_1 probability of '*Monday at* E' (that is, of '*somewhere* E, and *everywhere*, $E \supset$ *Monday*') is zero. By the same reasoning, you should be certain that it is not Tuesday. On the other hand, you should be certain that it is either Monday or Tuesday! P_2 ends up not being a probability distribution at all. (It does not help to redefine '*Monday at* E' as '*somewhere*, *Monday* & E'; then you should become certain that it is both Monday and Tuesday.)

Now given the richness of our evidence, worlds where evidence propositions are true more than once are strange. Can we dismiss the trouble cases as far-fetched and content ourselves with a rule that works at least in normal

situations (following [Titelbaum 2008] and [Stalnaker 2008: ch.3])? I don't think so. For one, the nature of our evidence is a contingent matter. It is easy to imagine (or build) creatures whose evidence is rather sparse. More importantly, the contradiction arises as soon as worlds with multiple occurrence of the evidence are assigned non-zero credence. But how far-fetched is it to assign non-zero credence e.g. to the hypothesis of eternal recurrence, on which history keeps repeating itself? Indeed, statistical mechanics arguably entails that the universe, if large enough, contains many short-lived duplicates of our present brains, emerged from random collisions of atoms in outer space and time (see [Albrecht and Sorbo 2004]); if evidence supervenes on brain states, these so-called *Boltzmann brains* have the same evidence that we have. Similarly, Everettian quantum mechanics arguably entails that many of our counterparts in other branches of the universe have the same evidence that we have. It does not matter whether these theories are *true*. What matters is that situations where someone assigns them *non-zero credence* can hardly be dismissed as far-fetched.

To allow for such cases, one may appeal to a principle of *self-locating indifference*: if the present evidence is true at several points within a world, these points should all get equal credence. (IC) might then be extended as follows. As before, start by ruling out all worlds from the previous belief space where E occurs nowhere. To re-introduce the centers, divide the probability assigned to worlds where E occurs more than once evenly between all the E locations. [Halpern 2006] considers something like this proposal. In effect, we re-interpret '$P_1(A \textit{ at } E)$' in (IC) as the expected ratio of A locations among E locations. Another possibility, suggested in [Piccione and Rubinstein 1997] and [Elga 2000], is to redistribute probability from worlds with fewer occurrences of E to worlds with more of them, by multiplying the previous probability of each world with the number of E locations it contains, dividing its probability evenly between these locations, and renormalising the probability function.

Either way, the appeal to indifference leads to a new host of problems. What if the number of possible centers within a world is infinite? Given a world of (one-way) eternal recurrence, what is the "indifferent" probability for living in an even-numbered epoch, or in a prime-numbered epoch? Any simple answer to such questions quickly runs into contradiction. Worse, indifference is the high road to skepticism. If you believe in statistical mechanics and hence that the

universe may well contain legions of Boltzmann brains all of whom share your present evidence, self-locating indifference requires you to be confident that you are one of these brains. (Elga [2004] appears to welcome this consequence.)

Shifted conditioning has none of these problems. It gives the correct verdict in cases like *Eternal Life*, it easily handles multiple occurrences of the evidence, and it makes no use of indifference.[7]

6 Sleeping Beauty and diachronic rationality

Many recent proposals on updating and self-location have been developed in response to the Sleeping Beauty problem. A quick reminder:

> **Sleeping Beauty.** While Sleeping Beauty is asleep on Sunday night, a fair coin is tossed. If the coin lands tails, Beauty's memories of Monday will be erased on Monday night. If the coin lands heads, her memories aren't erased, but she is made to sleep all through Tuesday. Beauty knows of this arrangement before she goes to bed on Sunday.

The "problem" is what Beauty should believe about the outcome of the coin toss when she wakes up on Monday.

There are two reasons why I haven't discussed this case yet. One is its strangeness. Self-location spells trouble for conditioning in simple, everyday situations like *The Litmus Test*; we don't need to look at far-fetched puzzle cases that no-one ever encounters. More importantly, while it is pretty clear what rationality demands in *The Litmus Test* or *Eternal Life*, there is considerable disagreement about *Sleeping Beauty*. This makes *Sleeping Beauty* rather ill-suited as a starting point or test case for a general model.

The other reason why I have so far ignored *Sleeping Beauty* is that it combines two issues that are better kept apart: self-location and memory loss (or

7 A problem somewhat parallel to multiple occurrences of the evidence arises for shifted conditioning if there can be multiple "next" points where information arrives. As it stands, shifting is undefined in such cases. Whether such cases are possible depends on somewhat controversial assumptions in the epistemology of fissioning subjects. Personally, I think they are possible, and I have worked on an extension of the present framework to deal with them. Due to space constraints, I have to leave this for another occasion.

"imperfect recall"). I will get back to this in a minute. First let's see what happens if Beauty follows (SC).

If Beauty follows (SC), her Monday credence in *Heads* equals her Sunday credence in $\succ$*Heads* conditional on $\succ E$, where E is her evidence on Monday. Assuming that her Sunday credence in *Heads* was 1/2 and $\succ E$ is irrelevant to the outcome of the coin toss, $P_2(\mathit{Heads}) = P_1(\succ \mathit{Heads} \mid \succ E) = 1/2$. But this is not the traditional "halfer" solution. On Sunday, Beauty was also confident that her next awakening would take place on Monday. Hence upon awakening, Beauty is certain that it is Monday! Stranger still, since she is now certain that her next awakening will take place on a Tuesday, she will be certain on Tuesday that it is Tuesday!

You may wonder whether it is even possible for Beauty to satisfy these demands. Imagine upon awakening on Monday morning, she wants to apply (SC). To this end, she first has to figure out her previous beliefs. Once she finds out that last night she believed that it is Sunday, she can infer that it is now Monday. But how is she supposed to find out what she believed last night unless she already knows that it is Monday? Indeed, if the coin lands tails, then Beauty may well have the exact same evidence on Monday and on Tuesday; and then this evidence can hardly lead her to one conclusion on Monday, and to a different conclusion on Tuesday.

This worry rests on an *evidentialist* picture of rational belief. Evidentialism is the doctrine that rational belief is never constrained by contingent matters outside the agent's present evidence (compare [Feldman and Conee 1985]). Hence if Beauty has the same evidence on Monday and on Tuesday, rationality cannot require her to have different beliefs.

Evidentialism has striking consequences for update policies. The policies considered so far all determine the new beliefs from two factors: the previous beliefs and the new evidence. But unless the previous beliefs are somehow part of the new evidence (in which case they are redundant as an additional factor), this puts constraints on the new beliefs that go beyond the agent's new evidence. An *evidentialist policy* would have to determine the new belief state entirely on the basis of the new evidence, drawing on previous beliefs only to the extent that they are recoverable from present evidence. Formally, such a policy is still a function from previous beliefs P_1 and evidence E to new beliefs P_2, but it is a degenerate function that completely ignores the input P_1.

The issue about evidentialism is somewhat muddled by the fact that people mean all kinds of things by 'evidence'. If you count anything as evidence to which a belief state is rationally sensitive, then evidentialism is true by definition. Obviously, (SC) cannot clash with a definition. To bring out the anti-evidentialist consequences of (SC), 'evidence' should be used for something that might uncontroversially be the same between Beauty's state on Monday and her state on Tuesday – something like sensory input, or sensory input together with accessible memory.

Returning to the above worry, is it possible for agents to follow update policies that are not evidentialist in this sense? It is. Consider a simple robot with a camera and a database for storing information. When the camera delivers new information, the information is added to the database. The robot's update mechanism thereby implements a non-evidentialist policy: the content of the database is determined not only by the present evidence from the camera, but also by what was written in the database before. To implement an evidentialist policy, the robot would have to erase the database every time it receives new information; the database content would always reflect just the present input from the camera. (If you think the previous content of the database should count as part of the robot's evidence, let me stipulate that the database is write-only: there is no way for the robot to read its contents.)

As the example illustrates, non-evidentialist update mechanisms have a great advantage: they allow agents to maintain beliefs even when they are no longer supported by present evidence. To the extent that the initial beliefs were well supported, this typically leaves the agent with a more accurate representation of her environment. Except in very unusual circumstances, (SC) easily outperforms any evidentialist rival in terms of truth-tracking. This might explain why no engineer would dream of implementing an evidentialist mechanism in a robot, and why nature almost certainly hasn't implemented one in us.

However, the worry about (SC)'s verdict on *Sleeping Beauty* might go deeper. What if Beauty's "memory erasure" affects not only her accessible memory, but every aspect of her mental state? What if Beauty's Tuesday morning state is an exact copy of her Monday morning state if the coin lands tails? It is then quite plausible (though not inevitable) that Beauty's beliefs on Monday cannot be different from her beliefs on Tuesday. So she cannot satisfy the demands of (SC).

This may be right, but it does not count against (SC) as a norm of diachronic rationality. In a cognitive system that reliably implements (SC), the new belief state must be causally sensitive to the previous state. If this causal sensitivity is destroyed and the system is brought into a state completely unrelated to its predecessors, it is only to be expected that the system will violate norms of diachronic rationality.

What lies at the heart of this matter is *whether there are any diachronic norms of rationality at all*, norms that directly relate an agent's beliefs at one time to her beliefs at another time. Evidentialism presumably rules out such norms: according to evidentialism, rational belief may be constrained by present evidence about previous belief, but it is unconstrained by previous belief itself. Any belief state may be followed by any other, as long as the new state fits the new evidence.

None of this has anything in particular to do with (SC). All the update norms I have considered in this paper are genuinely diachronic. Conditioning, for example, says that the agent's new credence should equal her previous credence conditional on the new evidence. It does not say that the new credence should equal *what the agent takes to be her previous credence, or what she can in some way recall as her previous credence*, conditional on the new evidence. Conditioning relates the agent's new credence to her actual previous credence, not to any present trace of the previous credence.

I find it very plausible that there are diachronic norms of rationality, but I have no hopes of proving this. If you disagree, you will easily find any argument to the contrary invalid or question-begging – a fate that has befallen every argument for conditioning. Trying to establish normative claims from uncontested grounds is a futile exercise.

Nevertheless, I have something on offer even if you reject diachronic constraints on belief, and thus any form of conditioning. For you may still like a corresponding principle of inverse Reflection. That is, you may agree that if an agent *knows* that her previous self believed, say, that yew berries are poisonous, then it is sensible for her to retain this belief, even without more direct evidence for its content. The agent could thereby partake in the *doxastic conservatism* that characterises followers of conditioning: she would tend not to change her

mind on a subject matter unless she encounters relevant evidence.[8]

To handle centered propositions, I would recommend the shifted version of inverse Reflection from section 3:

$$(\text{S}\text{Я}) \quad P_2(A) = \sum_x P_2(P_1(\succ A \,|\, \succ E) = x) \times x.$$

This says that your credence in A should equal your estimate (formally, your expectation) of your previous credence in $\succ A$ conditional on $\succ E$, where E is your present evidence. If this estimate coincides with the actual previous credence – as it does when you possess complete and certain evidence about your previous beliefs – then obeying (SЯ) has the same effect as following (SC), and it enjoys the same advantages over the Reflection principles corresponding to (C) and (IC).

Let me emphasise again that two independent issues are in play here. One is whether there are genuinely diachronic constraints on rational belief. If not, rules like conditioning or shifted conditioning can only be normative for agents with perfect recall. For situations with imperfect recall, we should retreat to the corresponding principle of inverse Reflection, (Я) or (SЯ). The second issue is how conditioning should be modified to handle centered information. This may look like a moot point if you reject diachronic norms of rationality anyway. But the problems for conditioning also affect the principle of Reflection: if propositions can change their truth-value, (Я) is just as unacceptable as (C).

My focus in this paper has been on the second issue. I have argued for (SC) as an alternative to (C), and correspondingly for (SЯ) as an alternative to (Я). In the case of *Sleeping Beauty*, (SC) says that Beauty should be confident on Monday that it is Monday and on Tuesday that it is Tuesday – even if this is impossible. 'Ought' does not always imply 'can'. On the other hand, if we set

8 This form of conservatism is only distantly related to its more prominent namesakes, discussed e.g. in [Christensen 1994] and [Vahid 2004]. It says nothing about the justificatory status of beliefs. By recommending conservative policies, I do not claim that unjustified beliefs can become justified merely by being retained. Nor is it a mark of conservatism, as I use it, that one always regards the fact that one used to believe p as evidence for p. This is impossible: let q be a proposition that entails that you used to believe $\neg q$ (such as a typical skeptical scenario); the conditional probability of $\neg q$ given that you used to believe $\neg q$ then cannot exceed the unconditional probability of $\neg q$; hence the fact that you used to believe $\neg q$ cannot possibly be evidence for $\neg q$. More generally, on my usage, a conservative agent need not have any evidence or opinions about her previous beliefs at all.

aside diachronic norms, or norms that are impossible for an agent to obey, we may look to the synchronic norm (SЯ). (SЯ) does not settle the answer to the Sleeping Beauty problem. However, if we assume that Beauty is not certain on Monday that it is Monday, (SЯ) entails that her credence in *Heads* should be less than 1/2. Combined with a principle of self-locating indifference, (SЯ) yields the traditional "thirder" solution.[9]

7 Conclusion

Let me wrap up. I have proposed a modified form of conditioning as a general rule for updating centered beliefs. Like conditioning, the rule is conservative: by default, old beliefs are carried over to the new state; a belief is dropped (loosely speaking) only if it either conflicts with the new evidence or was expected to become false due to changes in the world. The second clause distinguishes the modified rule, shifted conditioning, from standard conditioning.

When centered propositions are ignored, shifted conditioning reduces to conditioning. More specifically, the two coincide whenever $P_1(A \mid E) = P_1(\succ A \mid \succ E)$: when the present probability of A conditional on E equals the probability of A being true in the near future given that E is true in the near future.

Like conditioning, shifted conditioning has a synchronic counterpart (a principle of inverse Reflection) that imposes no diachronic constraints on belief. For agents with perfect information about their previous beliefs, the two coincide. Otherwise agents who follow shifted conditioning will usually end up with a more accurate representation of their environment than agents who merely obey the synchronic counterpart.

9 By the law of total probability, $P(H) = P(H \mid \mathit{Mon}) \times P(\mathit{Mon}) + P(H \mid \mathit{Tue}) \times P(\mathit{Tue})$. If it is Monday, then the previous probability of *Heads* was 1/2, and so was the previous probability that *Heads* will be true conditional on $\succ E$. Hence by (SЯ), $P(H \mid \mathit{Mon}) = 1/2$. Since $P(H \mid \mathit{Tue}) = 0$, the probability of *Heads* is $1/2 \times P(\mathit{Mon})$. This line of reasoning parallels the argument against the halfer solution in [Titelbaum 2008] and [Kim 2009], where the Reflection principle corresponding to (IC) is used in place of (SЯ).

I have assumed that Beauty's Tails-Tuesday state is a successor of her Tails-Monday state. Given the "memory erasure", it might be preferable to regard the Tails-Tuesday state as a direct successor of Beauty's Sunday state. The Sunday state would then have *two* successors, one on Monday and one on Tuesday. With the adjustments mentioned in footnote 7, both (SC) and (SЯ) then lead to the traditional "halfer" solution.

Unlike many proposals in the AGM tradition (following [Katsuno and Mendelzon 1991]), my account makes no assumptions about how the world will evolve. It is not assumed, for example, that the most probable future is exactly like the present. Unlike many proposals in the Bayesian tradition (such as [Halpern 2006], [Titelbaum 2008], [Meacham 2008] or [Kim 2009]), my account assigns no special status to uncentered beliefs or uncentered propositions. Uncentered beliefs are simply beliefs with a particular subject matter; a subject matter that does not distinguish between locations within any possible universe.

I have argued that shifted conditioning gives the correct verdict in everyday cases like *The Litmus Test* and *The Prisoner* as well as in cases like *Eternal Life* that pose problems for alternative proposals. I have also shown how several well-known arguments in favour of conditioning support shifted conditioning once centered propositions are taken into account.

I do not claim that shifted conditioning is the only rational way to change one's mind. It might be better to magically align one's beliefs with the truth, irrespective of the evidence and the previous beliefs. And I suppose there should be room for rationally revising one's beliefs without receiving any relevant evidence, as when one finds a new theory that better explains the available data.

In these respects, shifted conditioning shares whatever burden lies on standard conditioning; the only improvement is that it adequately handles centered propositions. I think of shifted conditioning as an *ideal conservative* policy – a conservative policy that does not lose track of changes in the world.

References

Albrecht A, Sorbo L (2004) Can the universe afford inflation? Physical Review D 70:063,528

Arntzenius F (2003) Some problems for conditionalization and reflection. Journal of Philosophy 100:356–370

Christensen D (1994) Conservatism in epistemology. Noûs 28(1):69–89

Diaconis P, Zabell SL (1982) Updating subjective probability. Journal of the American Statistical Association 77:822–830

Elga A (2000) Self-locating belief and the sleeping beauty problem. Analysis 60:143–147

Elga A (2004) Defeating dr. evil with self-locating belief. Philosophy and Phenomenological Research 69:383–396

Feldman R, Conee E (1985) Evidentialism. Philosophical Studies 48(1)

van Fraassen B (1984) Belief and the will. Journal of Philosophy 81(5):235–256

van Fraassen B (1999) Conditionalization, a new argument for. Topoi 18(2)

Goldstein M (1983) The prevision of a prevision. Journal of the American Statistical Association 78:817–819

Halpern J (2006) Sleeping beauty reconsidered: conditioning and reflection in asynchronous systems. In: Gendler T, Hawthorne J (eds) Oxford Studies in Epistemology, Vol.1, Oxford University Press, pp 111–142

Jeffrey R (1983) The Logic of Decision, 2nd edn. University of Chicago Press, Chicago

Katsuno H, Mendelzon A (1991) On the difference between updating a knowledge database and revising it. Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR-92) pp 387–394

Kim N (2009) Sleeping beauty and shifted jeffrey conditionalization. Synthese 168:295–312

Kumar PR, Varaiya P (1986) Stochastic Systems. Prentice-Hall, Englewood Cliffs, NJ

LaValle SM (2006) Planning Algorithms. Cambridge University Press, Cambridge

Lewis D (1979) Attitudes *De Dicto* and *De Se*. The Philosophical Review 88:513–543

Lewis D (1999) Why conditionalize? In: Papers in Metaphysics and Epistemology, Cambridge University Press, Cambridge, pp 403–407

Meacham C (2008) Sleeping beauty and the dynamics of de se beliefs. Philosophical Studies 138:245–269

Meacham C (2010) Unravelling the tangled web: Continuity, internalism, non-uniqueness and self-locating beliefs. In: Gendler TS, Hawthorne J (eds) Oxford Studies in Epistemology, Volume 3, Oxford University Press, pp 86–125

Perry J (1979) The problem of the essential indexical. Noûs 13:3–21

Piccione M, Rubinstein A (1997) On the interpretation of decision problems with imperfect recall. Games and Economic Behavior 20:3–24

Stalnaker R (2008) Our Knowledge of the Internal World. Oxford University Press, Oxford

Teller P (1973) Conditionalization and observation. Synthese 26(2)

Titelbaum MG (2008) The relevance of self-locating beliefs. The Philosophical Review 117:555–606

Vahid H (2004) Varieties of epistemic conservatism. Synthese 141(1)

Williams PM (1980) Bayesian conditionalisation and the principle of minimum information. British Journal for the Philosophy of Science 31(2):131–144