

# Changing minds in a changing world

Wolfgang Schwarz <wo@umsu.de>

Draft, May 26, 2009

**Abstract.** I propose an update rule for belief that takes into account both the impact of new evidence and changes in the subject's location. I argue that this rule can play the same role in a centered worlds framework that conditioning plays in the classical possible worlds framework. I also discuss whether we need diachronic rationality constraints at all, how my proposal relates to other recent proposals, and what results we get for puzzles like the Sleeping Beauty problem.

## 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. Believing that tomorrow is Tuesday is not the same thing as believing tomorrow that it is Tuesday. If you were to ask me today whether it is Tuesday, I would say no; tomorrow, I would say yes. At some point in between, I will stop believing that it is Monday, and start believing that it is Tuesday.

Philosophers disagree on how to model this kind of belief change. The most straightforward treatment says that apart from information about the universe as a whole, there is also information 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 as a whole, but for *me, now*: it might have been 23:50, or 23:58.

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’ without qualification, 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 by conditioning on new evidence, the resulting beliefs are typically incompatible with the old ones. These two kinds of belief change have been extensively studied in stochastic control theory and the AGM school of formal epistemology; 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.<sup>1</sup>

I will present an update rule that incorporates both kinds of belief change, loosely building on ideas from control theory. Even though I work in the centered-worlds framework, my proposal should also be applicable to other accounts. Thus [Perry 1979] and others have argued that belief should be understood as a three-place relation between a subject, an uncentered content, and a centered *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 (see [Chalmers 2008]). 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 describes the dynamics of these diagonal probabilities.

## 2 Diachronic rationality

In the centered-worlds framework, a belief state is modelled by a probability distribution over centered worlds. To simplify the presentation, I will pretend that the class of worlds is finite; nothing essential hinges on this. When new evidence comes in, the probability distribution may be revised. The classical rule for such revisions,

---

<sup>1</sup> See [LaValle 2006: part III] for an overview of stochastic control theory (also known as sequential decision theory) with further references. [Katsuno and Mendelzon 1991] introduces the two kinds of update in the AGM tradition; see [Boutilier 1998] for an AGM-style account closely related to control theory and thereby to the ideas developed in the present paper.

*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  $E$ :

$$P_2(A) = P_1(A | E). \tag{C}$$

Different rules have been proposed, for instance to allow for evidence that does not raise the probability of any proposition to 1. I will briefly return to these in section 8.

What counts as ‘evidence’ in this context? Formally it does not matter. For concreteness, I will occasionally follow philosophical tradition and talk as if the relevant kind of evidence was somehow tied to experience, so that, for example, the fact that there is smoke in your kitchen is not enough for you to have evidence that your food is burning; you also have smell or see or otherwise notice the smoke. Likewise for information stored in your brain: it does not count as evidence unless you are aware of it. The main arguments for my proposal do not presuppose this or any other specific conception of evidence.

Let an *update policy* be a function that takes a probability distribution and an evidence constraint as input and returns a new probability distribution as output. 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 the constraint imposed by her total new evidence. Until section 8, I assume that evidence constraints always say that some evidence proposition  $E$  has probability 1. An update policy can then also be specified as a function from a probability distribution  $P_1$  and a proposition  $E$  to a new probability distribution  $P_2$ .

The conditioning policy (C) makes direct use of the previous credence. It does not state that the agent’s later credence equals *what she takes to be her previous credence* conditional on her new evidence. Agents who follow conditioning may well have no idea about their previous beliefs. To be sure, a cognitive mechanism that reliably implements conditioning would have to be sensitive to the previous credence; but this need not affect the agent’s evidence. Our brain manages to regulate our blood oxygen level even though we rarely have specific evidence about the present level.

An update policy like (C) should therefore not be understood as a recipe for how to deliberately set your credence based on the currently available evidence. Without evidence about your previous beliefs, you cannot just decide to follow conditioning. Think of the choice between update policies as a choice in engineering: what kind of mechanism would you choose if you were to build a robot, or if you were to implement (by neurosurgery) a policy in yourself? More to the point, what kind of policy would

best advance the goals of epistemic rationality?

Some policies, let's call them *purely evidential*, determine the new beliefs entirely on the basis of the new evidence, drawing on previous beliefs only to the extent that they are recoverable from the present evidence. Conditioning is not a purely evidential policy, but it has an "internalised" counterpart that might figure in a purely evidential policy:

$$P_2(A | P_1(A | E) = x) = x. \quad (C^*)$$

This says that upon receiving evidence  $E$ , your new credence in any proposition  $A$  equals your estimate (formally, your expectation) of your previous conditional credence in  $A$  given  $E$ . 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  $(C^*)$  has the same effect as following  $(C)$ . If not, not.

Very different mechanisms are also conceivable. Instead of a mechanism that operates on the agent's previous beliefs, one could have a mechanism that operates on her *future* beliefs, or on the previous beliefs of the agent's neighbour (compare [Christensen 2000]). Yet other mechanisms would directly align the agent's beliefs with the facts, without any detour through evidence. Those mechanisms do not implement update policies on my definition, since they do not determine the new credence based on the old credence and the new evidence alone. By focusing on mechanisms that are in this sense *local*, I do not mean to suggest that non-local alternatives are necessarily worse; perhaps they are just much harder to implement.

Purely evidential policies are noteworthy because they impose no direct diachronic constraints on beliefs.  $(C^*)$  merely constrains the relation between present beliefs and a certain subset of them concerning previous beliefs. The *actual* previous beliefs never enter the picture. Like conditioning, the policy I will defend draws directly on the agent's previous beliefs, whether or not they can be recovered from present evidence. It thereby clashes with strongly "evidentialist" views according to which rational belief should only be constrained by facts of which the agent is presently aware.

Trying to defend rationality constraints from uncontested ground is a hopeless enterprise. Philosophers who reject diachronic rationality constraints will easily find any argument to the contrary invalid or question-begging – a fate that has befallen every argument for conditioning. Some of these arguments will be reviewed in section 8; for now, I leave it with the following observation. It has been suggested that the ultimate epistemic goal is *true belief*, and that we value justification or support by the evidence only because well-supported beliefs tend to be true (see e.g. [Beckermann

2001], [Goldman 2001]). If this is on the right track, one can make a good case for diachronic rationality constraints: by following a rule like conditioning, agents can 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 than she would have if she followed a purely evidential policy on which beliefs have to be dropped as soon as they are no longer supported by present evidence. (See section 7 for examples.)

Be that as it may be, I have something on offer even if you reject diachronic constraints on rationality. For you might still be interested in their synchronic counterparts, in principles like (C\*). That is, you might think that at least if an agent is *aware* that her previous self believed, say, that yew berries are poisonous, then it might be sensible for her to retain that belief, even without more direct evidence for what it says. She could thereby partake in the *doxastic conservatism* that characterises followers of conditioning: she will tend not to change her mind on a subject matter unless she encounters relevant evidence.<sup>2</sup>

Perhaps conditioning and its synchronic counterpart are a bit *too* conservative. We should arguably allow for situations where agents reconsider their opinions without responding to new information. For agents who occasionally undergo memory corruption or acquire beliefs in irrational ways, it might also be useful to slowly discount beliefs that are never again supported by independent evidence, perhaps by mixing non-conservative priors into  $P_1$  before conditioning.<sup>3</sup>

There is another respect in which conditioning is too conservative: if an agent travels through space and time, it can be quite unreasonable for her to stick to her beliefs until contrary evidence comes along. When the church bell has struck midnight, I should not keep believing that it is midnight until I receive contrary information about the time. Even worse, policies like (C) tend to leave agents forever certain of their present

---

<sup>2</sup> 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  $\neg q$ . Moreover, on my usage, a conservative agent need not have any evidence or opinions about her previous beliefs at all.

<sup>3</sup> What gets conditioned on the new evidence would then be  $\kappa \times P_1 + (1 - \kappa) \times P_0$ , where  $P_0$  is a fixed “prior” distribution and  $\kappa$  a constant  $\in [0, 1]$ . The smaller  $\kappa$ , the less the system trusts its previous state and the more it privileges present evidence over past evidence.

evidence; I would be stuck with the enduring conviction that the church bell strikes midnight.<sup>4</sup>

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 these changes into account.

### 3 Indirect conditioning

To clarify what exactly is wrong with conditioning, consider another 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  $E$  be the information you receive, and  $P_1$  your credence function just before you see the paper turn red.  $P_1$  assigns high probability to worlds where the paper is about to turn red and somewhat lower probability to worlds where the paper will turn blue. Very low probability goes to worlds where the paper has already turned red; after all, you clearly see that the paper is white. Perhaps you cannot completely rule out worlds where the paper is red, or even worlds where the paper *looks* red. 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 notice that the paper looks red. Conditioning would have you move all your credence to worlds of the last kind, 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 onto your belief that either the paper is white or you are being fooled by evil demons.

A natural response is then to limit conservatism to uncentered matters. Fortunately, it is easy to extract the uncentered fragment from a centered credence function, as

---

<sup>4</sup> This problem for conditioning has often been noted, see e.g. [Oddie 1994: 466]. In the next section, I will present a version of the problem that does not presuppose that evidence makes agents certain of anything.

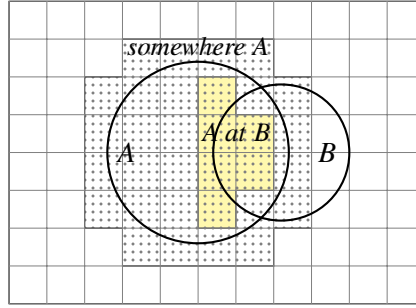


Figure 1: **uncentered propositions in centered logical space.** The grid partitions the space of worlds by the worldmate relation: two worlds are in the same cell iff they agree in what they say about the universe as a whole. ‘*Somewhere A*’ and ‘*A at B*’ express uncentered propositions about the universe,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

any centered proposition *A* determines an uncentered way the universe might be: to be such that *A* is true somewhere. For instance, the centered proposition that it is raining excludes any universe where it never rains anywhere; hence its uncentered fragment is the proposition that it rains at some time somewhere. Let ‘*somewhere A*’ express this uncentered fragment of *A*; it is uncentered insofar as it does not distinguish between different centers within the same universe. Let us also introduce ‘*A at B*’ as short for ‘*somewhere B*, and *everywhere, B ⊃ 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. (See figure 1.)

Now assume  $P_1$  gives zero credence to worlds where *E* is true more than once (in the history of that world). Then we could use the following policy of *indirect conditioning*:

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

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* obtains. In very broad outline, this is the policy recommended in [Halpern 2006], [Titelbaum 2008], [Kim 2009], and, though less explicitly, [Elga 2000] and [Elga 2004].

Consider *The Litmus Test*. Your total evidence  $E$  not only tells you that the paper is red, it also contains all kinds of information about the lab, the lighting, your memories, etc. Before receiving  $E$ , you were uncertain whether things would ever be just like that. Once you learn  $E$ , you know that you are in a universe where  $E$  obtains at least once. If you have independently ruled out any worlds where  $E$  obtains more than once, you can figure out your location within the  $E$  worlds: you must be at the very place where  $E$  obtains.

Here is another way to motivate this policy. 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 midnight on 12 April 2009 and I am alone in 53 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 any proposition:

$$P_2(A) = P_2(A \text{ at } L) = P_1(A \text{ at } L | E \text{ at } L). \quad (\text{IC2})$$

To apply (IC2), we need a suitable self-locating certainty  $L$ . Presumably  $L$  must come from the new evidence; at least it can’t have been learned by (IC2). But if  $E$  entails  $L$ , and possibilities with multiple occurrences of  $E$  are ruled out, then ‘ $E$  at  $L$ ’ reduces to ‘*somewhere*  $E$ ’; so (IC2) 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 (IC2) equivalent to (IC).<sup>5</sup>

What if worlds where  $E$  occurs more than once are not ruled out? 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, for 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.)

So (IC) collapses into contradiction when positive probability goes to worlds with more than one occurrence of  $E$ . Can we dismiss this as a far-fetched possibility and content ourselves with a rule that works at least in normal situations (following

---

<sup>5</sup> The versions of indirect conditioning proposed in [Titelbaum 2008] and [Kim 2009] resemble (IC2), those in [Halpern 2006] and [Elga 2000] look more like (IC).

[Titelbaum 2008] and [Stalnaker 2008: ch.3])? I don't think so. How far-fetched is it to assign positive probability to worlds of eternal recurrence, where each epoch contains people like us, with the exact same evidence? Worse, some respectable physical theories might entail that we are not alone with our present evidence. If the world is large enough, then statistical mechanics makes it very likely that at several points in space and time, various collections of atoms spontaneously form a temporary duplicate of our present brain (see [Albrecht and Sorbo 2004]). If evidence supervenes on brain states, these "Boltzmann brains" have exactly the same evidence that we have. Or think of no-collapse versions of quantum mechanics, on which we constantly branch into many successors, some of which presumably have the exact same evidence. 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.

A more far-fetched, but also more well-known example of such a predicament is the Sleeping Beauty problem.<sup>6</sup>

**Sleeping Beauty.** On Sunday, Sleeping Beauty is put to bed; then a fair coin is tossed. If the coin lands tails, someone will see to it that Beauty's awakenings on Monday and on Tuesday are completely indistinguishable. In particular, all memories she would normally have of Monday will be erased before she awakens on Tuesday. If the coin lands heads, nothing strange happens. Beauty knows of this arrangement.

The problem is what Beauty should believe when she wakes up on Monday. She knows that if the coin landed tails, then her present evidence  $E$  is true twice, once on Monday and once on Tuesday. Since she has no way to distinguish between the two cases, one might intuit that she ought to give them equal credence. More generally, one might intuit a principle of self-locating indifference: if the present evidence is true at several places within a world, they should all get equal credence.

A simple way of extending (IC) accordingly is this. 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

---

6 The problem was introduced somewhat abstractly as 'example 5' in [Piccione and Rubinstein 1997] and made popular in philosophy by [Elga 2000] and [Lewis 2001]. The present version, like that in [Piccione and Rubinstein 1997] and unlike that in [Elga 2000], does not stipulate that Beauty is asleep on Tuesday if the coin lands heads, while it does stipulate that the two awakenings are subjectively indistinguishable. Many prominent accounts, including [Elga 2000], [Halpern 2006], [Titelbaum 2008] and [Meacham 2008], but not the one I will defend, give different answers depending on how these details are filled in.

evenly between all the  $E$  locations. This is more or less Halpern's [2006] proposal. In effect, we re-interpret ' $P_1(A \text{ at } E)$ ' in (IC) as the expected ratio of  $A$  locations among  $E$  locations. For *Sleeping Beauty*, this leads to the "halfer" solution  $P_2(\text{Heads}) = 1/2$ ,  $P_2(\text{Tails \& Monday}) = P_2(\text{Tails \& Tuesday}) = 1/4$ .

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. This yields the "thirder" solution  $P_2(\text{Heads}) = P_2(\text{Tails \& Monday}) = P_2(\text{Tails \& Tuesday}) = 1/3$ .<sup>7</sup>

These accounts make indirect conditioning applicable to cases where the evidence occurs more than once. However, the appeal to indifference has curious consequences. Assuming that evidence supervenes on brain states, it entails that you should be confident that you are a Boltzmann brain if you believe that the universe is sufficiently large. Elga [2004] appears to welcome this consequence.

It gets worse. Suppose at  $t_1$  you assign low (but positive) credence to some no-collapse version of quantum mechanics. At  $t_2$  you see that dark clouds are gathering. On the no-collapse interpretation, your previous self has branched into many successors, some of which have exactly this evidence and others who do not. The number of people in the universe who share your present evidence is therefore significantly greater given the no-collapse theory than given its common alternatives. And their number is constantly increasing. On an Elga-style account, your credence in the no-collapse theory should therefore continuously rise as you watch the clouds.

The lesson is that we should not re-distribute probability from worlds with fewer occurrences of  $E$  to worlds with more of them. Halpern gets this right. But the problem returns in a different form. Suppose cloud formation is a chancy process. On the no-collapse theory, it was certain beforehand that on some branch the clouds would gather in exactly the way you find them. On the collapse theory, they might well have

---

<sup>7</sup> A further possibility is indicated, though not explicitly endorsed by [Kim 2009], who generalises (IC2) to

$$P_2(A) = \sum_i P_1(A \text{ at } L_i | E \text{ at } L_i) \times P_2(L_i),$$

where the  $L_i$  are the self-locating possibilities open at  $t_2$ . Kim does not say how to determine the  $P_2(L_i)$  if they are not given by the new evidence. Applying a principle of indifference with Monday and Tuesday as candidates, we get a "quarterer" distribution:  $P_2(\text{Heads}) = P_1(\text{Heads at Monday} | E \text{ at Monday}) \times 1/2 + P_1(\text{Heads at Tuesday} | E \text{ at Tuesday}) \times 1/2 = 1/2 \times 1/2 + 0 \times 1/2 = 1/4$ .

gathered only in different ways. In general, the no-collapse theory may well entail ‘*somewhere E*’ for any evidence proposition *E* we ever encounter. (IC) therefore only rules out collapse worlds, and your credence in the no-collapse theory should steadily rise.

All this has nothing in particular to do with quantum mechanics, nor even with multiple occurrences of the evidence. Consider this possibility.

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

Your current credence in *Eternal Life* is presumably low. But conditional on *Eternal Life*, the probability of any experience *E* occurring at some point or other is very high, whereas it is generally much lower on alternative hypotheses about your life. According to (IC), any experience whatsoever should therefore strongly raise your credence in *Eternal Life* over such alternatives.

The root of these problems is that indirect conditioning is entirely non-conservative 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 a moment ago you were certain that you live in the 21st century on Earth, (IC) will ignore this belief and try to figure out your new location from scratch, based on independent new evidence.

Followers of indirect conditioning are thereby not only led to curious beliefs, they also become vulnerable to Dutch Books. Suppose again that you assign credence  $x \in (0, 1)$  to *Eternal Life*, and suppose you know that you will not die in the immediate future. Then you look out of the window. By following (IC), your credence in *Eternal Life* is guaranteed to go up, no matter what you observe. If you bet in accordance with your beliefs, I can sell you a wager against *Eternal Life* now and buy it back at a reduced price later. (The general Dutch Book argument against indirect conditioning will be proved in section 8.)

I will now introduce a more conservative update policy that avoids these problems.

## 4 Shifted conditioning

I haven’t explained yet what the ‘previous credence function’ is that an update policy is meant to take as input. It can’t be an arbitrary credence from any earlier time. Any 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. It 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.) Of course, when we discuss particular cases, we may 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.

We 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; we should allow for agents who are not so fortunate as to always have complete evidence about everything they have ever learned.

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* onto 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 world and any time, there is a closest “next” time when (relevant) information arrives. How should rational credence evolve between these times? I suggest that if at  $t_1$  the agent believed that things are such-and-such, then in the absence of contrary evidence, she ought to believe at  $t_2$  that things were such-and-such just before the present evidence arrived. We do not hold fixed her self-locating beliefs; we do not assume that things are still exactly the way they were before. But nor do we completely ignore those beliefs. We assume – tentatively, and subject to revision in the light of new evidence – that the previous beliefs represent how things were a moment earlier.

To render this precise, let ‘ $>A$ ’ express the proposition that  $A$  will be true at the next point when evidence arrives. That is,  $>A$  is true at world  $w$  iff  $A$  is true at the next point when information arrives at  $w$ . The shifting operator  $>$  induces a transformation on the space of probability functions, mapping the credence  $P_1$  to the *shifted credence*  $P_1^>$ , with  $P_1^>(A) = P_1(>A)$ . Here, then, is the policy I recommend:  $P_2$  should equal  $P_1^>$

conditioned on the new evidence. Equivalently, the new credence in  $A$  upon learning  $E$  should equal the previous credence that  $A$  is true at the next point when evidence arrives, given that this evidence is  $E$ :

$$P_2(A) = P_1(>A | >E). \quad (\text{SC})$$

If  $A$  is certain not to change its truth-value in the foreseeable future, then  $P_1(>A) = P_1(A)$ . If both  $A$  and  $E$  have this property, then (SC) reduces to conditioning. (SC) might therefore be understood as a generalisation of conditioning, adjusted to handle centered propositions.

Evidence that is anticipated with certainty does not affect the probability of uncentered propositions. If at  $t_1$  you were certain that the next thing you'd learn is  $E$ , and also that  $A$  will not change its truth-value, then learning  $E$  at  $t_2$  will not affect your credence in  $A$ : if  $P_1(>E) = 1$ , then  $P_2(A) = P_1(>A | >E) = P_1(>A) = P_1(A)$ . This protects followers of (SC) from the kind of Dutch Book mentioned in the previous section.

There is an obvious resemblance between (SC) and (IC2). Both apply conditioning to a shifted transformation of the previous credence. (IC2) shifts the probability of any world  $w$  to wherever  $L$  holds within the same universe; (SC) shifts it to the next point at  $w$  when information comes in.<sup>8</sup>

Consider again *The Litmus Test*. Let  $t_1$  be the time when you start dipping the paper. Suppose at this point, 70% of your credence  $P_1$  goes to worlds where the paper is about to turn red, 29% to worlds where it is about to turn blue, and 1% to other ways things might be. Among worlds where the paper turns red, 99% of your credence goes to worlds where the liquid is acidic, reflecting your confidence in the test's reliability. Your *shifted* credence  $P_1^>$  therefore gives probability 0.7 to the paper being red, 0.29 to it being blue, and conditional probability 0.99 to the liquid being acidic given that the paper is red. Conditioning on your observation turns this conditional probability into an unconditional probability; you end up 99% confident that the liquid is an acid. No probability is moved to evil demon worlds.

Unlike indirect conditioning, shifted conditioning also gives the desired result in cases like *Eternal Life*. *Eternal Life*, remember, is the hypothesis that you will have any humanly possible experience at some point in your life. It does not say which of these experiences you will have next. Hence when you find that your next experience is  $E$ , this rules out just as many *Eternal Life* possibilities as normal possibilities. *Eternal*

---

<sup>8</sup> [Kim 2009] even calls his version of (IC2) "Shifted Strict Conditionalization". Understanding (SC) as an improved version of (IC2) may justify keeping the label.

*Life* is not strongly confirmed by every experience. Since shifted conditioning makes no use of indifference principles, it even allows us to believe that we are not Boltzmann brains.

## 5 Branching

Indirect conditioning ran into problem in cases where the evidence proposition  $E$  is true at more than one place in a world. Parallel problems would arise for shifted conditioning if at some worlds there are multiple “next points when information arrives”. Is this possible? Perhaps. Consider a case of fission.

**Fissioning Fred.** Fred’s home planet, *Sunday*, is surrounded by two moons, *Monday* and *Tuesday*. Tonight, while Fred is asleep, his body is scanned and destroyed; then a signal is sent to both Monday and Tuesday where he will be recreated from local matter. Fred is aware of all this.

Let  $P_1$  be Fred’s credence before falling asleep. What should his successor on Monday believe when he awakens, assuming his evidence is compatible with both Monday and Tuesday? It would not be reasonable to be confident that he does not exist, or that he is still on Sunday. Nor should he be confident that he is on Monday. (Any local update policy yielding this result would also make Fred’s successor on Tuesday confident that he is on Monday.) The right attitude, it seems, would be to give equal credence to being on Monday and being on Tuesday.

But if Fred updates his beliefs in line with (SC), then  $P_2$  can only assign credence 1/2 to ‘being on Monday’ if  $P_1$  already assigned credence 1/2 to ‘being on Monday at the next point when information comes in’. So the question is: could Fred have been genuinely *uncertain* whether he would wake up on Monday or on Tuesday when he went to bed on Sunday?

A number of philosophers have argued that he could (see [Saunders 1998], [Ismael 2003], [Wallace 2008], [Saunders and Wallace 2008]). For suppose a centered world is something like a triple of an uncentered world  $u$ , a time  $t$ , and a person  $i$ . Following [Lewis 1976], one can argue that there are *two* persons in *Fissioning Fred*: Fred<sub>M</sub>, waking up on Monday, and Fred<sub>T</sub>, waking up on Tuesday. The two have a common past, somewhat like two buildings with a common wall. When we point at Fred on Sunday, before the fissioning, we strictly speaking point at two people, Fred<sub>M</sub> and Fred<sub>T</sub>. So we have two centered worlds here,  $\langle u, t, \text{Fred}_M \rangle$  and  $\langle u, t, \text{Fred}_T \rangle$ . Moreover, at this stage Fred<sub>M</sub> arguably does not know that he is at  $\langle u, t, \text{Fred}_M \rangle$  rather than  $\langle u, t, \text{Fred}_T \rangle$ .

(He couldn't have learned it from evidence: whatever evidence he has is shared by Fred<sub>T</sub>.) So here we have the pre-fission ignorance we need:  $P_1$  assigns equal credence to Fred<sub>M</sub> possibilities where the next awakening takes place on Monday, and to Fred<sub>T</sub> possibilities where it takes place on Tuesday. Applying (SC) leaves Fred<sub>M</sub> undecided between Monday and Tuesday.

Let's call this account of fission *No Real Branching*. If No Real Branching is true, there can't be multiple next points where information arrives, not even in cases of fission. But No Real Branching is controversial (see [Greaves 2004], [Lewis 2007]). Imagine being in Fred's situation. It is certainly tempting to think that there are two ways things could be: either you will wake up on Monday or on Tuesday. But perhaps this is because it is tempting to suppose that there is something – a non-physical soul, a stream of consciousness – that settles all questions of survival, that gets transferred either to Monday or to Tuesday. But what if there is no such something? What if the whole truth is that you are about to fission into a Monday and a Tuesday successor, both of which are related to you in every way that normally matters for survival? [Parfit 1984] argues that the question who of them is *identical to you* is based on a mistaken metaphysics. If that is true, there may not be anything for pre-fission Fred to be uncertain about. On this view (*Real Branching*), worlds don't always have at most one next point when information arrives. In fission cases, we then do face the same problem that indirect conditioning had with multiple occurrences of the evidence.

To get clear about the options, return to *Fissioning Fred*, but assume that Fred is not sure on Sunday whether he will be teleported to both Monday and Tuesday or only to Monday. His credence is then divided between two kinds of worlds: worlds where he has only a Monday successor, and worlds where he has both a Monday and a Tuesday successor. Suppose he gives equal credence to both possibilities. How should these probabilities be distributed among his successors? (See figure 2a.)

One option is to distribute the credence of a branching world evenly among its successors. This amounts to redefining  $P_1(>A)$  as the expected ratio of  $A$  locations among the next points when information comes in, analogous to the Halpernian redefinition of ' $P_1(A \text{ at } E)$ '. Upon awakening, Fred would give credence 3/4 to being on Monday and 1/4 to being on Tuesday.

This looks plausible, but it might not be the best general account. If survival comes in degrees, as Parfit argues, perhaps the probability shifted to a successor world should reflect the degree of survival. Or perhaps the redistribution of credence should reflect the quantum mechanical amplitude of the respective branches, if the branching is due to spreading quantum entanglement (see [Greaves 2007]).

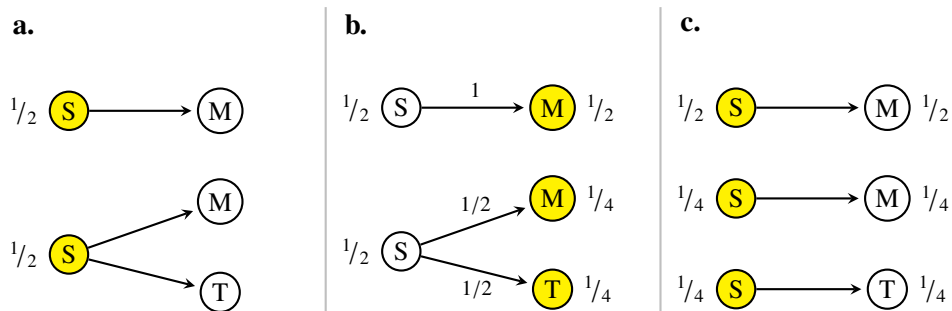


Figure 2: **shifting probabilities in branching worlds.** a. Fred is undecided between two possibilities: either he is about to be teleported to Monday, or he is about to be teleported to both Monday and Tuesday. b. Fred's credence is shifted to the successor probabilities, weighted by the transition probabilities. c. The No Real Branching account: the second possibility – being teleported to both Monday and Tuesday – is treated as two distinct possibilities from the outset.

A precise update policy would have to settle these matters. Here I prefer to remain neutral by simply associating *transition probabilities* with pairs of centered worlds. These are values that specify how a world's probability gets redistributed over its successors. I write  $P(v \succ w)$  for the transition probability between worlds  $v$  and  $w$ .<sup>9</sup> As before, the update process applies conditioning to a shifted probability  $P_1^\succ$ , but now  $P_1^\succ$  assigns to any world  $w$  the sum of the  $P_1$  value for worlds with  $w$  as a successor, weighted by the transition probability between the two (see figure 2b).

<sup>9</sup> Transition probabilities are a familiar parameter in stochastic control theory, but they play a slightly different role there. In control theory, subjective probability is defined not over centered worlds, but over fragments of stages of worlds, called *states*. States don't 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 about how these states will evolve. This is represented by the transition probabilities. (In most treatments, transition probabilities are actually derived values, determined by the probability of various *events* or *nature actions*, and the probability that those events will lead to the relevant outcome.) Replacing the incomplete states with complete worlds, we usually have several possibilities where control theory says there is only one: if state  $v$  could develop into  $w_1$  or  $w_2$ , depending on whether nature does this or that, then we have two distinct worlds, one where nature does this and one where it does that. The worlds may be locally similar, but they are nonetheless different. The only place where we still need transition probabilities are cases of branching, where a single, maximally specific way things might be has multiple futures. (By using complete worlds, we also get rid of the Markov assumption: that the probability of future states depends only on the present state, irrespective of its history.)

It will be useful to extend transition probabilities to arbitrary propositions (rather than single worlds). To this end, let  $P_1(B \succ A)$  be the expectation of arriving at an  $A$  world from a  $B$  world, weighted by the probability of the  $B$  worlds; that is,

$$P_1(B \succ A) := \sum_w \sum_{v \in A} P_1(w | B) \times P(w \succ v).$$

If for  $B$ , we plug in the tautologous class of all worlds (denoted by the empty symbol ‘ $\succ$ ’), we get the probability of arriving at an  $A$  world from wherever we might be now. This is of course just the shifted probability  $P_1^{\succ}(A)$  of  $A$ . That is,  $P_1^{\succ}(A) = P_1(\succ A)$ . If we also define  $P_1(\succ A | \succ E)$  as  $P_1(\succ (A \& E)) / P_1(\succ E)$ , we can express shifted conditioning the same way as before:

$$P_2(A) = P_1(\succ A | \succ E). \tag{SC}$$

Shifting never moves probability between branching and non-branching possibilities. If Fred assigned credence 1/2 to non-branching possibilities on Sunday, his new credence in not having undergone branching is still 1/2. We could alternatively have assigned to each successor world the *full* probability of its predecessor and then renormalised the shifted probability, mirroring Elga’s proposal concerning worlds with multiple  $E$ s. But this would have led to the same unfortunate consequences: information about the weather would come out as strong evidence that the universe is constantly branching.

What if  $v$  has *no* next point when information arrives, because the agent faces oblivion? If we let  $P(v \succ w)$  be zero for all  $w$ , then  $P_1^{\succ}$  may sometimes fail to be a standard probability function, although  $P_2$  still is. We could alternatively treat terminal worlds as succeeded by a designated *null world* that is always excluded by evidence. I see no grounds for choosing between these options. (The choice would matter if describing an update policy was describing an implementation. But update policies are just functions from an earlier probability and an evidence constraint to a later probability. When I speak of ‘first’ shifting and ‘then’ conditioning, this is how I prefer to calculate the function; the two steps are not meant to have any psychological reality.)

Let me emphasise that transition probabilities are not some new kind of primitive probability. Here is how you determine  $P(v \succ w)$ . First, ask if  $v$  has a unique next point when information arrives. If yes,  $P(v \succ w)$  is either zero or one, depending on whether  $w$  is that next point. On No Real Branching, this covers all cases, and we’re done. If you prefer Real Branching, you have to decide how probability should be distributed if  $v$  has multiple “next points”. If you think that the probability should be evenly divided

among successors, then  $P(v \succ w) = 1/n$ , where  $n$  is the number of  $v$  successors (or zero if  $w$  is not among the successors). If  $v$  is an Everett world and you think the redistribution of probabilities should follow the quantum mechanical amplitudes (see e.g. [Wallace 2007]), then  $P(v \succ w)$  is the squared modulus of the amplitude of the  $w$  branch diverging from  $v$ . And so on for other trouble cases.

Advocates of No Real Branching cannot really avoid these questions. Where Real Branchers see a single possibility with multiple futures, they see multiple possibilities with unique futures. Since these possibilities are at present indistinguishable, the question arises how rational credence should be divided among them: evenly, or in accordance with the quantum mechanical amplitudes, or ...? For any given answer, the result of applying (SC) will be exactly the same as on the corresponding specification of transition probabilities for Real Branchers (see figure 2c). The dispute between the two camps is therefore largely irrelevant to our discussion, and I will ignore it from now on. Note that on No Real Branching, the revised version of (SC) reduces to the original version. Hence I take the revised version to be the official version, although its definition is somewhat roundabout from the perspective of No Real Branching.

I will sometimes make use of transition probabilities between unspecific propositions like *Monday* or *Tails*. These have the same value on either account of branching. To illustrate, let *Tails* be the proposition that Fred is teleported to both Monday and Tuesday, and *Heads* the proposition that he is only teleported to Monday. On No Real Branching, *Tails & Sunday* divides into two sub-cases, one where the unique successor is on Monday, and one where he is on Tuesday. If the two cases have equal probability,  $P_1(\textit{Tails \& Sunday} \succ \textit{Tails \& Monday})$  is  $1/2$ , just as it is on Real Branching if the transition probabilities are equal.

We can picture the update process in a table.

		H Mon	T Mon	T Tue
<i>1/2</i>	H Sun	1	0	0
<i>1/2</i>	T Sun	0	1/2	1/2
Shifting:		<i>1/2</i>	<i>1/4</i>	<i>1/4</i>
Conditioning:		<i>1/2</i>	<i>1/4</i>	<i>1/4</i>

The body of the table gives the transition probabilities between the relevant (unspecific) propositions. The italicised values at the left are the old probabilities, the values at the bottom the new ones. The ‘Shifting’ values are calculated by adding up the numbers in each column multiplied by the old probability of the relevant row. The conditioning step is idle here since we have stipulated that nothing relevant is learnt when the two Freds awaken.

If we skip conditioning and apply shifting to the shifted probabilities, we get the double-shifted probability  $P_1^{2>}$ , the triple-shifted probability  $P_1^{3>}$ , etc. Intuitively,  $P_1^{n>}(A)$  is the  $t_1$  credence that  $A$  will be the case after  $n$  intakes of new information. Let us call this proposition (that  $A$  will be the case after  $n$  intakes of new information) ‘ $>_n A$ ’. So  $P_1^{n>}(A) = P_1(>_n A)$ . As we will see next,  $P_1(>_n A)$  is also the expectation of your future credence in  $A$ .<sup>10</sup>

## 6 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*:

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

Informally, to satisfy Reflection means to trust one’s (expected) future judgement. This kind of trust is evidently silly if propositions can change their truth-value. I suppose that tomorrow I will believe that it is Tuesday; this does not mean that I should already believe today that it is Tuesday. If we allow for centered propositions, 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*:

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

Shifted Reflection characterises self-aware agents who follow (SC). (Proof: let  $E_1, \dots, E_n$  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  $P_1(>A | >E) = P_1(>A | >E')$ . For each  $i \leq n$ , let  $x_i = P_1(>A | >E_i)$ , where  $E$  is any member of  $E_i$ . By the law of total probability,  $P_1(>A) =$

<sup>10</sup> Algebraically, the shifted probability vector  $P_1^{>}$  results from multiplying the transition probability matrix  $\mathbf{T} : \mathbf{T}_{i,j} = P_1(w_j > w_i)$  with the probability vector  $P_1$ . (Compared to the table above, rows and columns in  $\mathbf{T}$  are swapped.) The  $n$ -shifted probability  $P_1^{n>}$  is the result of  $n$  matrix multiplications:  $P_1^{n>}(w) = (\mathbf{T}^n P_1)(w)$ .

$\sum_i P_1(> E_i) \times x_i$ . If you know that you update by (SC),  $P_1(> E_i \leftrightarrow P_2(A) = x_i) = 1$ . Hence  $P_1(> A) = \sum_i P_1(P_2(A) = x_i) \times x_i$ , and you satisfy (SR).

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. For instance, 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 is 1. In general, (SR) only holds if  $P_2$  is presented as 'my credence at the next point when information comes in'. That is, we should read (SR) like this:  $P_1(> A) = \sum_x P_1(> P(A) = x) \times x$ . More generally, using the  $n$ -shifted probability function from the previous section,

$$P_1(>_n A) = \sum_x P_1(>_n P(A) = x) \times x. \tag{SR}$$

A nice illustration of these issues is Frank Arntzenius's [2003] story of the prisoner. We 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 time still knows what time it is. She also knows that each sleep phase takes either one or two hours. Her initial credence is distributed at follows.

<b>8pm</b>	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*.

<b>1st awak.</b>	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. Moreover, the combination *Execution & 12am* is excluded by the evidence that the lights are still on. By (SC), the new credence is<sup>11</sup>

<b>2nd awak.</b>	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:

<b>3rd awak.</b>	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 4th 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  $n$ th 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 that time. The expectation of her 11pm credence is therefore greater than 1/2. In general, whatever her credence in *Banishment* is at 8pm,

<sup>11</sup> Here is the update table:

		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

the expectation of her 11pm credence is higher (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 probability that the lights will be off by the second awakening is 1/8, the expectation of the future credence in *Banishment* is  $4/7 \times 7/8 = 1/2$ .<sup>12</sup>

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 align your beliefs with their unconditional judgement, but with their (expected) *conditional* judgement, conditional on the further information you possess. And 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.

## 7 Evidence, conservatism, and Sleeping Beauty

Reflection relates the present credence to the expected later credence. We can also relate it to the earlier credence. In this case, it is very common that we have relevant information that our previous selves lacked. Among self-aware agents who follow conditioning, the later credence therefore matches the expectation of the previous credence conditional on the new evidence. More specifically,

$$P_2(A | P_1(A | E) = x) = x. \quad (C^*)$$

We have seen this before: it is the "synchronic counterpart" of conditioning mentioned in section 2. The other policies I have discussed also have synchronic counter-

---

<sup>12</sup> 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 that 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 evidence 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.)

parts. For (IC), it is

$$P_2(A | P_1(A \text{ at } E \mid \text{somewhere } E) = x) = x. \quad (\text{IC}^*)$$

For (SC), we get

$$P_2(A | P_1(>A \mid >E) = x) = x. \quad (\text{SC}^*)$$

It is easy to verify that self-aware agents who follow (C), (IC) or (SC) satisfy the corresponding synchronic principles. For example, suppose you follow (SC) and have just received the information  $E$ . Your new credence  $x$  in  $A$  then equals your previous credence in  $>A$  conditional on  $>E$ . Being aware of both your present credence and your update policy, you can conclude that your previous credence in  $>A$  conditional on  $>E$  was  $x$ . So  $P_2(P_1(>A \mid >E) = x) = 1$  and  $P_2(A | P_1(>A \mid >E) = x) = P_2(A) = x$ . (As before, we have to be careful that  $P_1$  is picked out in the right way: as your credence just before the present information came in.)

As I mentioned in section 2, the converse is false: satisfying one of the synchronic principles is not enough to follow the corresponding update policy. This is obvious as the starred principles only relate the present credence to present beliefs about the previous credence, irrespective of what the previous credence really was. Nevertheless, the synchronic and diachronic principles have the same result if the agent has complete and certain evidence about her previous beliefs.

An extreme violation of this condition occurs in the following story, again from [Arntzenius 2003].

**Shangri La.** There are two paths to Shangri La: the path by the mountains and the path by the sea. A fair coin is tossed to decide which path you will take; heads you go by the mountains, tails you go by the sea. If you go by the mountains, nothing strange happens. But if you go by the sea, then as soon as you enter Shangri La, your memories of the journey will be erased and replaced with (quasi-)memories of a journey by the mountains. The coin land heads.

As you travel past the mountains, you have no reason to distrust your experience; you are certain that you travel by the mountains. But what happens when you arrive in Shangri La? On the diachronic policies we have considered – (C), (IC), and (SC) – you retain this belief; you remain certain that you went by the mountains.

This may seem strange. How can you be certain that you went by the mountains if you know that you would have the exact same evidence if you had come by the

sea? Is it not irrational to trust the evidence of your memories once you have arrived? It is. But on the diachronic policies, your belief is not just a matter of your present evidence. If you had travelled by the sea and followed one of these policies, you would now be certain that you came by the sea, despite all the vivid memories of a mountain journey.<sup>13</sup> Since you know that you would have these memories no matter what, they are neutral on the question which path you took; hence a conservative policy will simply retain your previous beliefs on that matter. In *Shangri La*, all diachronic policies we have considered yield this conservative verdict. Obviously, none of their synchronic counterparts do so.

Note how the conservative policies outperform the synchronic versions in terms of truth-tracking: if you follow a conservative policy, you can retain potentially useful information about your journey that you would lose if you merely obey their synchronic counterparts. Imagine you build a robot that repeatedly travels to Shangri La. And imagine you would like to know which path the robot took on each occasion. Would you not implement an update policy that lets the robot retain this information, rather than one on which all information about the journey gets erased as soon as the robot encounters evidence *known to be irrelevant* to this question?

Still, the diachronic policies violate the doctrine that rational belief should be constrained only by present evidence, and not by “external” things like previous beliefs. For philosophers who accept only synchronic rationality constraints, I would recommend (SC\*). At least in situations where agents have perfect knowledge about their previous beliefs, it yields the same result as (SC), and enjoys the same advantages over (C\*) and (IC\*).

I will list some of these advantages next, after a final look at Sleeping Beauty. Recall that on Monday morning, Beauty remembers that a fair coin was tossed on Sunday, and that she has indistinguishable Monday and Tuesday awakenings iff that coin landed tails. By itself, this does not tell her whether it is Monday or Tuesday, or even Wednesday. One might think that given no further information, she ought to give equal credence to the four combinations of *Heads/Tails* with *Monday/Tuesday*. But Beauty has further information: the absence of memories from later than Sunday. This rules out the *Heads & Tuesday* combination, as well as any possibility later than Tuesday. Beauty is left with credence 1/3 for the remaining three combinations. We get the thirder solution,

---

<sup>13</sup> Here I assume that what gets replaced after the journey by the sea is your memory experience. Things get more complicated if not only your experiences, but also your beliefs are overwritten upon entering Shangri La, or if the “memory erasure” somehow blocks the access to your previous beliefs.

defended roughly along the lines of [Horgan 2004].<sup>14</sup>

But what if Beauty's credence is not entirely determined by her present evidence? What if she follows a more conservative policy on which previous beliefs can be preserved without constant support by evidence? What if she follows (SC)? Her Monday credence in *Heads* then equals her Sunday credence in  $> \textit{Heads}$  conditional on  $> E$ , where  $E$  is her evidence on Monday. If her Sunday credence in *Heads* was  $1/2$  and  $> E$  is irrelevant to the outcome of the coin toss,  $P_2(\textit{Heads}) = P_1(> \textit{Heads} | > E) = 1/2$ .

But this is not the traditional halfer solution as defended in [Lewis 2001]. On Sunday, Beauty was also confident that her next awakening would take place on Monday. Hence upon awakening, (SC) makes her certain that it is Monday – despite the fact that her evidence would be just the same on Tuesday if the coin landed tails. In this respect, *Sleeping Beauty* looks much like *Shangri La*.

There is, however, another interpretation of the story on which we get the traditional halfer distribution. Imagine being the experimenter in *Sleeping Beauty*. How do you ensure on tails that Beauty will not be able to figure out that it is Tuesday when she awakens that day? You have to erase every trace Monday left on her. If she drank on Monday, she must not have a hangover; if she broke her wrist, she must not feel pain; if she learnt to speak Arapaho, she must have lost that ability. In effect, you have to undo everything that happened to Beauty on Monday, putting her back into the state she was after falling asleep on Sunday.

But now her situation looks a lot like that of Fissioning Fred, whose body was scanned on Sunday and recreated from local matter on Tuesday. (In Beauty's case, the recreation is made easier by the fact that the local matter – Sleeping Beauty in her post-Monday state – is already arranged not too different from the target arrangement.) If Tuesday-Fred counts as a direct successor of Sunday-Fred, then Tails-Tuesday-Beauty should count as a direct successor of Sunday-Beauty.

Suppose then that we interpret *Sleeping Beauty* as a case of branching. Under (SC), her credence evolves as follows (just like Fred's).

		H Mon	T Mon	T Tue
$1/2$	H Sun	1	0	0
$1/2$	T Sun	0	$1/2$	$1/2$
Shifting:		$1/2$	$1/4$	$1/4$
Conditioning:		$1/2$	$1/4$	$1/4$

<sup>14</sup> Since Beauty is not asleep all through Heads-Tuesday in my version of the story, we can bypass the objection to Horgan raised in [Pust 2008].

Once again, the conditioning step is idle, since the waking experience does not provide any unexpected information about the relevant propositions. Beauty ends up halving, and this time assigns credence  $3/4$  to *Monday* and  $1/4$  to *Tuesday*. If she later learns that it is Monday, her credence in *Heads* rises to  $2/3$ .

On this version of the story, Beauty knows that the immediate predecessor of her present state is her state on Sunday. Hence if she remembers her Sunday credence, we do not have to rely on genuinely diachronic principles. (SC\*) is enough to yield the traditional halfer distribution.

[Piccione and Rubinstein 1997], [Dorr 2002] and [Arntzenius 2003] independently came up with the following variation of *Sleeping Beauty* that is supposed to undermine this solution. Suppose Beauty’s memories are erased both on heads and on tails; but if the coin lands heads, she gets strong evidence for *Heads & Tuesday* immediately after waking up on Tuesday. When she wakes up on Monday, she ought to be indifferent at first between *Heads* and *Tails*; not finding the *Heads & Tuesday* evidence then should make her lean towards *Tails*. This is correct even on (SC), where the Monday update now goes

		H Mon	H Tue	T Mon	T Tue
$1/2$	H Sun	$1/2$	$1/2$	0	0
$1/2$	T Sun	0	0	$1/2$	$1/2$
Shifting:		$1/4$	$1/4$	$1/4$	$1/4$
Conditioning:		$1/3$	0	$1/3$	$1/3$

Whence the difference? Beauty’s perceptual evidence on Monday may be the same in either story. In either case, she can rule out *Heads & Tuesday*, and none of the other three possibilities. However, in the original story, Beauty knew on Sunday that if the coin lands heads, then her only successor will awaken on Monday. In the modified story, the “memory erasure” ensures that she also has a successor on Tuesday. For conservative agents, this difference in previous beliefs makes a difference to the later beliefs.<sup>15</sup>

## 8 Conditioning revisited

Many arguments have been given in support of conditioning, showing that under various circumstances, it is the only update rule that satisfies certain desiderata. Once we allow propositions to change their truth-value, it turns out that these arguments

<sup>15</sup> For parallel reasons, (SC) does not yield halving in [Bostrom 2007]’s variation *Beauty the High Roller*.

actually support shifted conditioning. I will illustrate this with three well-known, and hopefully representative, examples: an argument from coherence, an argument from Reflection, and an argument from minimal revision.

Before we start, let me briefly explain how shifted conditioning can be generalised to other revision rules and evidence constraints. So far I have assumed that the effect of new evidence is to raise the probability of some proposition  $E$  to 1. Richard Jeffrey [1983: 164–169] argues that we should use a weaker constraint on which the new evidence merely determines a distribution of probabilities  $x_1, \dots, x_n$  over a partition of evidence propositions  $E_1, \dots, E_n$ . His rule of *generalised conditioning* then says that

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

Since shifted conditioning is a combination of shifting and conditioning, it is straightforward to replace the conditioning step with generalised conditioning:

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

Similar adjustments are possible for other variations of conditioning.

Now on to the coherence argument, due to [Lewis 1999].<sup>16</sup> 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  $P^{-E}$  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 that  $E$  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 benefit is  $P_1(E)$  times that value. 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 your present

---

<sup>16</sup>Lewis's argument was first presented in [Teller 1973]. [Armendt 1980] provides a version for generalised conditioning. Here I will focus on the simple form.

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 *current* probability of  $E$  and  $A$ . If these propositions 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 hung it out recently, 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 actual expected payoff of the arrangement is  $-x \times P_1(>(E \& \neg A)) + (1 - x) \times P_1(>(E \& A))$ , which is greater than zero iff  $x < P_1(>(E \& A))/P_1(>E)$ . Thus what the coherence argument really shows is that your updated credence  $P^E(A)$  should equal your conditional *shifted* credence  $P_1(>A | >E)$ .

The next argument I want to consider 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

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

As [van Fraassen 1999] shows, if an agent satisfies Reflection, 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 we plug in centered propositions, but Reflection itself becomes implausible. (You should not believe that it is Tuesday merely because you think that this is what you will believe tomorrow.) As I argued in section 6, trusting one's future self is better expressed by a shifted version of Reflection,

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

(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  $w$ ,  $P_1(> w) = P_1(> E) \times P^E(w) + P_1(> \neg E) \times P^{\neg E}(w)$ . If  $w \models E$ , then  $P^{\neg E}(w) = 0$  and  $P_1(> w) = P_1(> E) \times P^E(w)$ . Moreover,  $P_1(> w) = P_1(> E) \times P_1(> (w \& E)) / P_1(> E)$ ; hence  $P^E(w) = P_1(> (w \& E)) / P_1(> E) = P_1(> w | > E)$ . On the other hand, if  $w \models \neg E$ , then  $P^E(w) = 0$  and  $P_1(> w | > E) = 0$ , so again  $P^E(w) = P_1(> w | > E)$ . So  $P^E$  results from  $P_1$  by (SC).

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

$$\text{if } A \text{ and } B \text{ entail } E, \text{ and } P_1(A) = P_1(B), \text{ then } P^E(A) = P^E(B). \quad (\text{MR})$$

[Teller 1973: 225–232] proves that conditioning is the only general rule for  $P^E$  that satisfies this constraint.<sup>17</sup>

The minimal revision principle is in line with conservatism: it entails that in the absence of relevant evidence, credence does not change. But 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 that  $E$ , you have found strong evidence against  $A$ , and not against  $B$ . We should therefore replace (MR) by a shifted version:

$$\text{If } A \text{ and } B \text{ entail } E, \text{ and } P_1(> A) = P_1(> B), \text{ then } P^E(A) = P^E(B). \quad (\text{SMR})$$

For uncentered  $A$  and  $E$ , (SMR) reduces to (MR), just as (SR) reduces to (R). It is easy to see that shifted conditioning satisfies (SMR): if  $A$  and  $B$  entail  $E$ , then  $P_1(> (A \& E)) = P_1(> A)$  and  $P_1(> (B \& E)) = P_1(> B)$ , hence if  $P_1(> A) = P_1(> B)$ , then  $P_1(> (A \& E)) / P_1(> E) = P_1(> (B \& E)) / P_1(> E)$ . The converse, that *no other* policy satisfies (SMR), can be proved by straightforward but tedious adaptation of the proof in [Teller 1973]; I leave this as an exercise to the reader.

## 9 Conclusion

Let me wrap up. I have proposed a modified form of conditioning as a general rule for updating centered beliefs. Like conditioning, my rule is conservative: by default, old beliefs are carried over to the new state; a belief is dropped (loosely speaking) if it

---

<sup>17</sup> See [Diaconis and Zabell 1982] and [Williams 1980] for related results. [Joyce 2004: 148f.] presents this as the main justification for conditioning.

either conflicts with the new evidence or was expected to become false due to changes in the world. The second clause distinguishes the revised rule, shifted conditioning, from classical conditioning.

When centered propositions are ignored (as they often are), shifted conditioning reduces to conditioning. More specifically, the two coincide whenever  $P_1(A|E) = P_1(\succ A|\succ E)$ : 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 purely evidential counterpart that imposes no diachronic constraints on belief. For agents with perfect evidence 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.

Unlike many proposals in the AGM/KM 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 my account gives more sensible verdicts than others in certain cases involving branching and evidence propositions that are true at several points in a universe. I have also shown how well-known arguments that have traditionally been taken to support conditioning actually 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. It might also be better to discount old evidence as mentioned in section 2. And I suppose there should be room for rationally revising one's beliefs without receiving any new 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 (local) conservative* policy – a conservative policy that does not lose track of changes in the world.

Thanks to Jens Christian Bjerring, David Chalmers, John Cusbert, Kenny Easwaran, Alan Hájek, Stephan Leuenberger, Tobias Rosefeldt, Weng Hong Tang, Michael Titelbaum and J. Robert G. Williams for helpful comments on earlier drafts.

## References

- Andreas Albrecht and Lorenzo Sorbo [2004]: “Can the universe afford inflation?” *Physical Review D*, 70: 063528
- Brad Armendt [1980]: “Is there a dutch book argument for probability kinematics?” *Philosophy of Science*, 47(4): 583–588
- Frank Arntzenius [2003]: “Some problems for conditionalization and reflection”. *Journal of Philosophy*, 100: 356–370
- Ansgar Beckermann [2001]: “Zur Inkohärenz und Irrelevanz des Wissensbegriffs. Plädoyer für eine neue Agenda in der Erkenntnistheorie”. *Zeitschrift für Philosophische Forschung*, 55: 571–593
- Nick Bostrom [2007]: “Sleeping beauty and self-location: A hybrid model”. *Synthese*, 157: 59–78
- Craig Boutilier [1998]: “A unified model of qualitative belief change: a dynamical systems perspective”. *Artificial Intelligence*, 98: 281–316
- David Chalmers [2008]: “Probability and Propositions”. URL <http://consc.net/papers/probability.pdf>. Forthcoming
- David Christensen [1994]: “Conservatism in epistemology”. *Noûs*, 28(1): 69–89
- [2000]: “Diachronic coherence versus epistemic impartiality”. *Philosophical Review*, 109(3): 349–371
- Persi Diaconis and Sandy L. Zabell [1982]: “Updating Subjective Probability”. *Journal of the American Statistical Association*, 77: 822–830
- Cian Dorr [2002]: “Sleeping Beauty: In defence of Elga”. *Analysis*, 62: 292–296
- Adam Elga [2000]: “Self-locating belief and the Sleeping Beauty problem”. *Analysis*, 60: 143–147

- [2004]: “Defeating Dr. Evil with Self-Locating Belief”. *Philosophy and Phenomenological Research*, 69: 383–396
- Alvin Goldman [2001]: “The Unity of Epistemic Virtues”. In Abrol Fairweather and Linda Zagzebski (Eds.) *Virtue Epistemology. Essays on Epistemic Virtue and Responsibility*, Oxford: Oxford University Press, 30–48
- Michael Goldstein [1983]: “The Prevision of a Prevision”. *Journal of the American Statistical Association*, 78: 817–819
- Hilary Greaves [2004]: “Understanding Deutsch’s probability in a deterministic multiverse”. *Studies in History and Philosophy of Modern Physics*, 35: 423–456
- [2007]: “On the Everettian epistemic problem”. *Studies in History and Philosophy of Modern Physics*, 38: 120–152
- 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
- Terry Horgan [2004]: “Sleeping Beauty Awakened: New Odds at the Dawn of the New Day”. *Analysis*, 64: 10–21
- Jenann Ismael [2003]: “How to Combine Chance and Determinism: Thinking about the Future in an Everett Universe”. *Philosophy of Science*, 70: 776–790
- Richard Jeffrey [1983]: *The Logic of Decision*. Chicago: University of Chicago Press, 2 edition
- James Joyce [2004]: “Bayesianism”. In Alfred Mele and Piers Rawling (Eds.) *The Oxford Handbook of Rationality*, New York: Oxford University Press, 132–155
- H. Katsuno and A.O. Mendelzon [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)*: 387–394
- Namjoong Kim [2009]: “Sleeping Beauty and Shifted Jeffrey Conditionalization”. *Synthese*, 168: 295–312
- Steven M. LaValle [2006]: *Planning Algorithms*. Cambridge: Cambridge University Press

- David Lewis [1976]: “Survival and Identity”. In Amelie O. Rorty (Hg.), *The Identities of Persons*, University of California Press, 17–40
- [1979]: “Attitudes *De Dicto* and *De Se*”. *The Philosophical Review*, 88: 513–543. Reprinted in Lewis’s *Philosophical Papers*, Vol. 1, 1983.
- [1999]: “Why Conditionalize?” In *Papers in Metaphysics and Epistemology*, Cambridge: Cambridge University Press, 403–407
- [2001]: “Sleeping Beauty: Reply to Elga”. *Analysis*, 61: 171–176
- Peter Lewis [2007]: “Uncertainty and Probability for Branching Selves”. *Studies in History and Philosophy of Modern Physics*, 38: 1–14
- Christopher Meacham [2008]: “Sleeping Beauty and the Dynamics of De Se Beliefs”. *Philosophical Studies*, 138: 245–269
- Graham Oddie [1994]: “Harmony, Purity, Truth”. *Mind*, 103: 451–472
- Derek Parfit [1984]: *Reasons and Persons*. Oxford: Clarendon Press
- John Perry [1979]: “The problem of the essential indexical”. *Noûs*, 13: 3–21
- 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
- Simon Saunders [1998]: “Time, Quantum Mechanics, and Probability”. *Synthese*, 114: 373–404
- Simon Saunders and David Wallace [2008]: “Branching and Uncertainty”. *British Journal for the Philosophy of Science*, 59: 293–305
- Robert Stalnaker [2008]: *Our Knowledge of the Internal World*. Oxford: Oxford University Press
- Paul Teller [1973]: “Conditionalization and observation”. *Synthese*, 26(2)
- Michael G. Titelbaum [2008]: “The Relevance of Self-Locating Beliefs”. *The Philosophical Review*, 117: 555–606
- Hamid Vahid [2004]: “Varieties of epistemic conservatism”. *Synthese*, 141(1)

Bas van Fraassen [1984]: “Belief and the will”. *Journal of Philosophy*, 81(5): 235–256

— [1999]: “Conditionalization, a new argument for”. *Topoi*, 18(2)

David Wallace [2007]: “Quantum Probability from Subjective Likelihood: improving on Deutsch’s proof of the probability rule”. *Studies in History and Philosophy of Modern Physics*, 38: 311–332

— [2008]: “Epistemology Quantized: circumstances in which we should come to believe in the Everett interpretation”. *British Journal for the Philosophy of Science*, 57: 655–689

P. M. Williams [1980]: “Bayesian conditionalisation and the principle of minimum information”. *British Journal for the Philosophy of Science*, 31(2): 131–144