Diachronic norms for self-locating beliefs

Wolfgang Schwarz
Draft, 29 March 2017

Abstract. It is widely acknowledged that Bayesian conditionalization is not an adequate norm for updating self-locating beliefs, if these are modelled in terms of centred propositions. In response, some have suggested that we should not model self-locating beliefs by centred propositions, others have suggested that such beliefs are not subject to diachronic norms, and yet others have put forward a variety of new norms to take the place of conditionalization. I critically compare all these proposals and defend one particular answer.

1 Introduction

Two epistemic norms form the core of classical Bayesianism. The first, *probabilism*, is synchronic; it says that rational degrees of belief conform to the probability calculus. The second, *conditionalization* (or *Bayes' Rule*), is diachronic; it specifies how rational degrees of belief change as new evidence arrives. In the simplest case, where the new evidence is captured by a single proposition E that is learned with certainty, conditionalization says that the new credence $\mathrm{Cr}_{t+=}$ in any proposition A should equal the previous credence Cr_t conditional on E:

$$\mathrm{Cr}_{t+1}(A) = \mathrm{Cr}_t(A/E). \tag{C}$$

If the new evidence is more equivocal, determining new probabilities $x_1, \ldots, x_n$ over some evidence partition $E_1, \ldots, E_n$, (C) generalizes to (JC) (see [Jeffrey 1965]):

$$\mathrm{Cr}_{t+1}(A) = \sum_i \mathrm{Cr}_t(A/E_i)x_i. \tag{JC}$$

Many arguments have been given in support of these rules. For example, [Lewis 1999] (first reported in [Teller 1973]) and [Skyrms 1987] show that probabilistically coherent agents are vulnerable to diachronic Dutch Books if and only if they do not change their beliefs by conditionalization; [Leitgeb and Pettigrew 2010] show that conditionalization maximizes the expected accuracy of the new belief state relative to the old belief state; [Greaves and Wallace 2006] established a parallel result for epistemic utility in place of accuracy.

These results all assume that the relevant propositions (A and E, or A and $E_1, \ldots, E_n$) do not change their truth-value from one time to another. Consequently, they fail

to support conditionalization if rational credence is defined over *centred* propositions whose truth-value can vary within a possible world. Indeed, it is widely recognised that conditionalization then become inapplicable.

Different authors have drawn different lessons from this observation. Some have concluded that credences should always be construed as uncentred. Others have taken the problem as (further) evidence that there are no diachronic norms on rational credence at all. Yet others have suggested that diachronic norms must be restricted to uncentred propositions. I will review these proposals in section 2 and argue that they do not provide a fully satisfactory response.

Another type of response is to develop new rules for the dynamics of self-locating beliefs. In sections 3–6, I will look at the main rules that have been proposed in philosophy and theoretical computer science. In section 3, I consider the idea that updating rational credence should go by *imaging* rather than conditionalizing on the new evidence; I argue that this yields sensible results only in a very limited range of cases. In section 4, I look at a more promising approach on which updating involves a process of first *shifting* the centres of one's doxastically accessible worlds and then conditionalizing on the new evidence. Unfortunately, this account seems to presuppose that an agent's belief updates are perfectly synchronised with objective time. In sections 5 and 6, I consider three ways of modifying the account so as to avoid this assumption. I argue that two of the modifications are problematic, but the third one seems to work. In section 7 I return to an alternative approach I set aside as incomplete in section 2, and show how it may be completed, building on the resources developed in sections 3–6.

A caveat before we begin. The proposals I will discuss have been developed in different contexts, with different background assumptions, different notation, and a focus on different problems. Here I will abstract away from these differences, presenting the core of the relevant proposals as they bear on my topic, in my own notation. I will also ignore various subtleties in the discussed proposals that do not affect the points I will make. For the full picture, the reader is advised to consult the cited works.[1]

2 Self-location and conditionalization

To motivate the search for new norms on the dynamics of self-locating belief—and to clarify what exactly we are looking for—I want to briefly review why we should allow

1 The survey [Titelbaum 2016] may also be found useful, as it approaches the present topic from a slightly different angle—asking how agents should "coordinate" their credences at different times—and looks at some proposals in more depth. On the other hand, Titelbaum's survey does not consider the view I discuss in section 3, and treats all the proposals from sections 4–6 as one, for which [Kim 2009] is taken as representative. As I will explain in note 7 below, on the way I draw the lines it is doubtful whether [Kim 2009] even belongs to this family.

the objects of credence to be centred, and why such credences are subject to diachronic norms.

The Bayesian concept of credence is a technical term. To understand why the objects of credence should include centred propositions we therefore have to look at some of the roles credences are meant to play. One key role lies in the theory of rational choice. According to Bayesian decision theory, rational agents choose actions that maximize expected utility relative to their credences and utilities, which is why rational behaviour can be explained by the agent's beliefs and desires. But as Perry [1977] vividly pointed out, the actions we choose often depend not only on our beliefs about the world as a whole, but also on where we locate ourselves in space and time. You and I may agree that somebody is being attacked by a bear and on all other relevant propositions about the objective world, but if you locate yourself as the person under attack while I take myself to be an onlooker, our rational response will be very different. Similarly, we may take quite different actions if you believe that an important meeting starts in half an hour while I believe that it starts now, even if we both agree that it starts at noon. If credences are to play their standard role in guiding and explaining actions, these observations suggest that our credences are not exhausted by our views about the objective world: it also matters where we locate ourselves within the world.

Another key role for the concept of credence lies in Bayesian confirmation theory, which models how and to what extent hypotheses are supported by an agent's evidence. Here, too, allowing for centred hypotheses and centred evidence has proved useful—for example, when trying to understand how the evidence that things *around us* are such-and-such bears on the hypothesis that the universe contains many (or few) places where things are such-and-such (see e.g. [Bostrom et al. 2002], [Sebens and Carroll 2017], [Arntzenius and Dorr 2017]). Indeed, as Lewis [1979] pointed out, one can seemingly imagine an agent who knows all objective facts about the world from a God's eye perspective but is still ignorant about who and where she is in the world, and might learn so through further evidence.

The standard way to accommodate these phenomena, going back to [Lewis 1979], is to construe the objects of credence in such a way that their truth-value may vary not only from possible world to possible world, but also from time to time, from place to place, and from individual to individual. Suitable objects are not hard to find. Lewis suggests identifying the objects of credence with properties; a property is "true" relative to a given individual at a time and a world just in case the individual instantiates the property at the time and the world. Other authors use sentence types as objects of credence, drawing on the natural sense in which, for example, the English expression 'it is raining' is true at some times and places and false at others. Another popular idea is to construe objects of credences as sets of triples of an uncentred world, a time, and an individual; the set is "true" relative to an individual at a time and a world just in case the corresponding triple

is in the set. Yet another option is to take centred propositions as theoretical primitives, or to construe them as *states of affairs* in the tradition of [Plantinga 1974] and [Pollock 1984]: a state of affairs like *the sun shining* can plausibly obtain at some times and places and not at others.

For what follows, the difference between these proposals will not be important. As a neutral label, I will call any object in the domain of an agent's credence function a *(centred) proposition*. Classical, uncentred propositions can be treated as limit cases of centred propositions (as [Lewis 1979] explains).

Probability theory requires that the space of propositions is closed under conjunction, disjunction, and negation. To simplify the following discussion, I will make the slightly stronger assumption that the propositions form a complete atomic Boolean algebra, so that propositions can be identified with sets of atoms in the algebra, where an atom is a maximally consistent conjunctions of propositions. I will sometimes refer to these atoms as *centred worlds*.[2]

It is sometimes objected that the phenomena reviewed above are instances of Frege's puzzle which supposedly has been solved without departing from classical views about the nature of propositions. I disagree: the decision-theoretic reasons to postulate self-locating beliefs are not instances of Frege's puzzle and do not generalize to (other) *de re* beliefs. But it does not matter. Even if we grant the objection, we plausibly still need centred objects of credence. For example, suppose we follow [Salmon 1986] and explain Frege's puzzle by appealing to "guises" under which classical propositions are believed. To make credences mesh with Bayesian decision theory and conformation theory, we then have to identify the objects of credence with propositions-under-guises, since the guises can matter to confirmation and rational choice (see e.g. [Chalmers 2011]). To handle the above phenomena in a parallel fashion, we will have to make the guises centred. As a result, the objects of credence—propositions *in the present, stipulative sense*—will be centred. We can happily allow that "propositions" in some other sense (referents of 'that'-clauses, perhaps) are uncentred, but since our topic is the dynamics of credence, that other sense is not immediately relevant.

To be sure, one might still try to account for the reviewed phenomena in some other way, without making the objects of credence centred. The most extended defence of this strategy is due to Robert Stalnaker (see [Stalnaker 1981], [Stalnaker 2008: ch.3],

2 If propositions are construed as sentences or states of affairs (say), then they are of course not literally identical to sets of maximally consistent conjunctions of propositions. However, under plausible assumptions, the quotient algebra of the propositions under the relation of logical equivalence will still be a complete atomic Boolean algebra, and hence isomorphic to the power set algebra of the atoms. (Intuitively, since every proposition is logically equivalent to a disjunction of maximally consistent conjunctions, and since probability theory requires logically equivalent propositions to have the same probability, we can always use sets of maximally consistent conjunctions as proxies for propositions, interpreting the sets as disjunctions of their members.)

[Stalnaker 2014: ch.5], [Stalnaker 2016]), who suggests that we model an agent's doxastic state as comprising an uncentred credence function together with "links" representing where the agent locates herself relative to any uncentred world she deems possible.[3] Without exploring such alternatives any further, I will from now on assume that the objects of credence are centred.

As a consequence, we must proceed with caution when we think about relationships between the attitudes of agents who locate themselves at different places or times. For example, suppose you utter 'it is raining' and thereby express high credence in the centred proposition that it is raining.[4] Even if I trust your assertion, I may not come to assign high credence to the same centred proposition: if I believe that you are located 500 km to the North of me (we are talking on the phone), I will rather come to believe that it is raining 500 km to the North. So successful communication arguably can no longer be modelled simply as transfer of beliefs (see [Weber 2013]). Similarly, if I happen to believe that it is not raining wherever I am, this does not constitute a genuine disagreement between you and me. The intuitive concepts of *agreeing*, *disagreeing*, *having the same belief*, etc., can no longer be analysed simply in terms of assigning high credence to either the same or to incompatible propositions.[5]

Analogous problems arise for beliefs of the same agent at different times—which brings us back to conditionalization. Suppose soon after our conversation on the phone, you fall asleep; when you later wake up, you do not receive any new information about the weather. Nonetheless, your confidence in the centred proposition that it is raining should decrease, especially if you have reasons to believe that it wouldn't rain the whole day. You should still be confident that it *was* raining when you fell asleep, but you should be less confident that is *is* raining. It is hard to see how this change could come about by conditionalization. Conditionalization only reduces you credence in a proposition if the new evidence is relevant to that proposition, but by assumption you do not gain any new evidence about the weather: whatever evidence E you receive, your previous credence in *rain* conditional on E may well have been high; yet your new credence is low.

Some authors have taken the present difficulty as (further) support for the view that we should not postulate diachronic constraints on rational belief at all. According to

3 Stalnaker does not explain in detail how these states should change over time. It turns out that even though the credence function is uncentred, it can't evolve simply by conditionalization, since the underlying probability space frequently changes.

4 By 'the centred proposition that it is raining' I mean a centred proposition that is true at a place and a time and a possible world just in case it is raining at that place at that time in that world. I will use this shorthand form throughout the essay.

5 A popular objection to centring objects of belief is that this would yield false predictions about when two subjects agree/disagree or believe the same/different things (e.g. [Stalnaker 2008: 50], [Stalnaker 2014: 114], [Bradley 2013]). It is important to realize that the relevant predictions follow only under a naive and entirely optional analysis of these concepts.

these authors, what we should believe at any point of time is simply a matter of our evidence at that time; the earlier beliefs don't matter (see e.g. [Levi 1980], [Williamson 2000], [Christensen 2000], [Horgan 2008], [Meacham 2008], [Moss 2014], [Hedden 2015], [Meacham 2016], [Arntzenius and Dorr 2017]). That is, if E captures your present total evidence, then your credence should equal some *ur-prior* P conditional on E, where the ur-prior is something like a measure of evidential relevance, not your previous credence function.

At first glance, "ur-prior conditionalization" (as [Meacham 2016] calls it) seems to avoid the problem centred propositions pose for classical conditionalization. But on closer scrutiny, the problem is not so easily discharged. The issue turns on the understanding of 'evidence'. In classical Bayesianism (as employed, for example, in Artificial Intelligence), the new evidence E that is plugged into conditionalization can be understood as the information conveyed to the agent through her senses at the relevant time. Such a *sensory conception* of evidence would render ur-prior conditionalization utterly implausible: the vast majority of our beliefs are not supported by present sensory evidence, and that does not make them irrational. Ur-prior conditionalization therefore requires a different conception of evidence on which our total evidence includes information we acquired on earlier occasions. This is where the problem returns. Before you fell asleep, you knew that it is raining; upon awakening, your total evidence no longer includes the information that it is raining, but rather the information that it *was* raining. So your earlier evidence that it *is* raining somehow transforms into your new evidence that it *was* raining. What are the general rules for these transformations?

A parallel issue arises for a more conservative family of views on which there are normative constraints on how one's uncentred credences should evolve over time, but no such constraints for centred credences. According to the most popular way of developing this idea, your new credences are determined by first conditionalizing your previous uncentred credences (i.e., the credences over uncentred propositions) on the uncentred information provided by your new evidence and then using the self-locating part of your new evidence to determine your location (see [Piccione and Rubinstein 1997], [Halpern 2006], [Meacham 2008], [Titelbaum 2008], [Briggs 2010], [Titelbaum 2013]).[6] [Moss 2012]

6 The account developed in [Titelbaum 2008] and [Titelbaum 2013] looks superficially different, but has essentially the same structure. Titelbaum notes that we can often "translate" any given centred proposition into an uncentred proposition which is certain to have the same truth-value and therefore must have the same probability. For example, if at t you are certain that nobody else in the history of the universe ever had or ever will have the very same total phenomenal experience X that you have right now, then you can be certain that *it is raining* has the same truth-value as *it is raining at the unique time and place where someone has phenomenal experience* X. Assuming that conditionalization is suitable for uncentred propositions, Titelbaum suggests that your new credence in the translated propositions should equal your previous credence conditionalized on your uncentred new evidence. To determine the new probability of centred propositions, we then have to use the centred part of

defends a variant approach on which your new credences are determined by first setting aside the previous self-locating beliefs, then using a certain part of your new evidence (provided by your "sense of time") to determine your location, and finally conditionalizing on the remainder of your centred and uncentred evidence.[7]

Like ur-prior accounts, these accounts plausibly require a non-sensory conception of evidence. Consider an agent—a robot perhaps—whose senses provide only limited information about the arrangement and colour of mid-sized objects in its surroundings. Suppose in order to reach some goal, the agent has to pass through two rooms with identical interior; in the first she has to exit on the left, in the second on the right. Having entered the first room, the agent's previous uncentred beliefs together with her sensory evidence do not settle whether she is in the first room or in the second. But surely the agent should not become uncertain about where she is. If the proposals just reviewed are to apply to this case, the agent's evidence must therefore be construed so that it includes the information that she is in the first room, determined by some transformation from her previous belief that she is about to enter that room.

Until further notice, I will henceforth reserve the term 'evidence' for sensory evidence. My question is how an agent's credences over centred and uncentred propositions should change as time passes and new information arrives from the agent's senses.[8] There may be reasons for tackling the question by first defining a richer notion of evidence that includes information the agent received at earlier times, suitably transformed to take into

your evidence to find new translations between centred and uncentred propositions. For example, if your new evidence entails that you have phenomenal experience Y, your new credence in *it is raining* might equal your updated credence in *it is raining at the unique time and place where someone has* Y. The upshot is that the new credences are determined by conditionalizing the uncentred previous credences on the uncentred evidence and then re-introducing the centres based on the new evidence.

7 [Kim 2009] defends a similar account, but he does not say where the new beliefs about the time come from: whether they are determined by some combination of the previous uncentred beliefs and the new evidence—as in Moss's account—or whether the previous centred beliefs also play a role. I will have a little more to say on the proposals of Moss and Kim below (footnotes 7, 11, 13, 14, and 17).

8 Sensory evidence should not be conflated with what philosophers call the 'content' of perceptual experience, namely the conditions under which we would judge the experience to be veridical. There are difficult questions on how to understand the information provided to an agent through her senses. Since these questions are orthogonal to the problems posed by centred propositions, I want bracket to them for the present discussion.

One may also wonder what exactly counts as a sense. We should obviously allow for internal senses. Do memories count? I'd say that "occurrent memories" count, but potentially occurrent memories (like merely possible visual experiences) do not, and neither do mere episodes of retaining a belief—the sense in which I have remembered all day that Astana is the capital of Kazakhstan. One can imagine an agent with an internal sense organ that constantly scans her belief state and determines the new beliefs entirely on the basis of these scans and the input from other senses. For such an agent, the earlier beliefs revealed through the scans may count as sensory evidence. But that is an absurdly inefficient way to maintain and update one's beliefs. The account I will end up defending is neutral on what, how much, or how little an agent learns through her senses.

account the passage of time. But to my knowledge the relevant evidence transformations have never been spelled out, so the proposal is hard to assess directly. In section 7, I will explain how the gap can be filled.

Before I get there, I will look at accounts on which an agent's credences are determined directly by her earlier credences and her new (sensory) evidence, without appeal to a further conception of non-sensory evidence that is subject to diachronic norms of its own.

3 Imaging

In [Katsuno and Mendelzon 1991], Katsuno and Mendelzon distinguish two rules for updating a knowledge base. The rules are best introduced by an example.

> At t you are confident that a certain fruit basket contains either two apples or an apple and a banana; you give equal credence to both possibilities. At $t+1$ you learn E: that there is no banana in the basket.

You might treat the new information as a reason to revise your beliefs, concluding that there are two apples in the basket. But you might alternatively treat E as a message about how the world has changed, without revealing anything new about what was the case at t. In particular, you may come to know E by learning that someone went out to remove any bananas from the basket. In that case, you should not conclude that there are two apples in the basket. Rather, your new credence should be divided between the hypothesis that the basket contains one apple and the hypothesis that it contains two.

The first kind of belief change Katsuno and Mendelzon call *revision*, the second *update*. Updating, they suggest, is "bringing the knowledge base up to date when the world described by it changes". Working in the AGM tradition of [Alchourrón et al. 1985], they characterize the two processes by axioms for operations on a qualitative (non-probabilistic) knowledge base. The probabilistic analogs of these two operations (as explained e.g. in [Kern-Isberner 2001], [Walliser and Zwirn 2002]) are conditionalization and imaging.

Imaging was originally introduced in [Lewis 1976b] for a somewhat different purpose, and works as follows. Let P be a probability measure over some "worlds" W, let $f : W \times \mathcal{P}(W) \to W$ be a function that maps any world w and event E to the "most similar" world w' to w that satisfies E, relative to some fixed similarity ordering; finally, let $[[A]]^w$ denote the truth-value of A at w. Then the *image of P on E*, which I'll write as $P(\cdot//E)$ (with two dashes), is defined by

$$P(A//E) = \sum_{w \in W} P(w)[[A]]^{f(w,E)}.$$

Intuitively, imaging on E shifts the probability mass of every world to the "most similar" world that satisfies E.

So we have a first proposal as to how an agent's credences should evolve in order to keep track of a changing world: if E is the new evidence received at $t+1$, then the new credence should equal the previous credence imaged on E:

$$\mathrm{Cr}_{t+1}(A) = \mathrm{Cr}_t(A//E). \tag{I}$$

The rule can be generalized to cases where the evidence is equivocal or where there is more than one "most similar" world, but let's stick with the easy case.

I have put 'most similar' in scare quotes because the relevant function f need not track intuitive similarity. Instead, f should assign to w and E whatever state of the world would result from minimally changing w into a state where E is true.

To see all this in action, return to the fruit basket scenario. At t, your credence is divided between two (coarse-grained) centred worlds or world states: a two-apples-no-banana world and a one-apple-one-banana world. Learning that somebody intervened so as to make it true that there are no bananas in the basket, your credence in the one-apple-one-banana world is moved to a one-apple-no-banana world: the world resulting from the one-apple-one-banana world by making 'no bananas' true. Your credence in the two-apples-no-banana world remains in place, because that world already satisfies the 'no bananas' constraint.

The imaging account is popular in some quarters of theoretical computer science, but has gained almost no traction among philosophers (though see [Leitgeb 2016]). For good reasons.

A minor point first. It is not an a priori truth that removing a banana from a basket containing an apple and a banana leads to a state in which the basket contains just an apple. So either the imaging function f encodes contingent causal information about the world—in which case one would like to know how this information is updated over time—or the agent's probabilities must be defined over much more fine-grained states than usually assumed, so that a given world fixes not only how many fruits are in the basket but also what would happen under various interventions. Let's assume we have such a fine-grained representation.

The main problem with the imaging rule (and its non-probabilistic analog) is that it only applies under conditions that are almost never satisfied: when the information one receives about a change in the world reveals nothing about what the world was like before.[9] Suppose you learn not that there is no banana in the basket, but more specifically that a banana has been removed from the basket. This is information about a change, but it also reveals that the basket contained a banana. Upon receiving the information, you should become confident that the basket contains one apple and no banana, but imaging does not apply (and would not deliver that result). Here we need a different rule for "bringing the knowledge base up to date".

9 I am not the first to raise this complaint; see e.g. [Boutilier 1998] and [Lang 2007].

It is hard to think of any realistic case where the information one receives is entirely neutral on what the world was like before. The most promising examples involve decisions or commands. If you are certain all along that you could make E true and then decide to make E true, one might think that your decision carries no news about what the world was like before. However even that is not generally true. In Newcomb's Problem, for example, a decision to two-box indicates that the opaque box is empty. The epistemic impact of reaching the decision from a prior state of indecision is therefore not adequately modelled by imaging.

Even in the rare and unusual case where the information one receives sheds no light on what the world was like before, imaging can give wrong results because changes in the world are not always revealed to the agent. Suppose again that at t you believe that the fruit basket contains either two apples or an apple and a banana, but suppose you also believe that a fly is sitting on the basket. Then you instruct somebody to remove all bananas, and at $t+1$ you learn that the instruction has been carried out: there are no bananas in the basket. Imaging preserves your credence in scenarios that already satisfy the new information (no bananas). Thus it would leave you certain that if the basket contains two apples then there is a fly on the basket. But if you know that flies generally don't sit at the same place for long, you should not remain certain about this.

In sum, imaging does not provide a satisfactory answer to our question. It only applies in rare and unusual cases, and even then tends to give the wrong answers.

4 Shifting

The next approach I am going to discuss is the *de facto* standard in control theory and artificial intelligence (see e.g. [Russell and Norvig 2010: ch.15]).

We saw in the discussion of imaging accounts that in order to describe how an agent's beliefs should change over time, we need to consider not only her beliefs about the present state of the world, but also her beliefs about how this state may evolve, either as the result of her actions or by itself. In artificial intelligence, such beliefs are commonly represented by a *transition model* that is added to the agent's probability measure over world states. In the simplest case, the transition model defines a conditional probability over world states at any time $t+1$ conditional on the state at the previous time t.[10]

With the transition model in place, there is an obvious two-step process for computing the probabilities over the new world state at $t+1$ in the light of new evidence: first, the previous probability over states at time t is projected forward to $t+1$ by the transition model; then the result is conditionalized on the new evidence.

10 More sophisticated models allow for continuous time, take into account the states before t when predicting the state at $t+1$, and include a further argument for an agent's actions.

Unfortunately, presentations of this approach tend to remain unclear on points that are central to our present topic. For example, equation (15.5) on p.572 of [Russell and Norvig 2010]—which in this respect is representative of the entire literature—expresses the complete update by the following equation:

$$P(X_{t+1}/e_1 \dots e_{t+1}) = \alpha P(e_{t+1}/X_{t+1}) \sum_{x_t} P(X_{t+1}/x_t) P(x_t/e_1 \dots e_{t+1}) \qquad \text{(RN)}$$

Here, $P(X_{t+1}/e_1 \dots e_{t+1})$ is meant to be the agent's probability at time $t+1$ over possible world states X_{t+1}, after having received evidence $e_1 \dots e_{t+1}$ at times 1 through $t+1$; $P(e_{t+1}/X_{t+1})$ is the probability of the (sensory) evidence e_{t+1} given state X_{t+1} as specified by the agent's "sensor model"; $P(X_{t+1}/x_t)$ is the probability of X_{t+1} given a particular hypothesis x_t about the previous state, as specified by the transition model; $P(x_t/e_1 \dots e_{t+1})$ is the agent's previous probability over world states at time t; α is a normalizing constant (the denominator in the application of Bayes' Theorem to compute conditionalization).

Since we are interested in how subjective probabilities should evolve over time, it is advisable to use different labels for the probabilities at different times: Cr_{t+1} and Cr_t, rather than simply P. Dealing with ideal agents, we can assume that the agent's probability measure at any time incorporates the evidence they have received up to then, so there is no need to explicitly conditionalize these measures on the history of evidence; so we can write $\mathrm{Cr}_t(X_t)$ instead of $P(X_t/e_1 \dots e_t)$. Moreover, having added time indices to the probability measures, we should arguably remove them from the objects of probability, to make clear that Cr_t and Cr_{t+1} can assign probabilities to centred world states (like *there being two apples in the basket*), not just to uncentred propositions specifying the state of the world at a given time (*there being two apples in the basket at* t). This fits the informal discussion in [Russell and Norvig 2010] and elsewhere, and in any case is required for the present proposal to bear on our topic.

The transition model specifies how world states may change from one point of time to the next. We could represent this by a primitive binary probability measure Cr^T over centred propositions, so that $\mathrm{Cr}^T(A, B)$ is the probability of the next state satisfying B if the present state satisfies A. But it will be useful to make the content of the relevant attitudes more explicit. After all, what is captured by Cr^T are ordinary conditional beliefs about how the world may change: about whether B is going to be true given that A is true now.

To model these beliefs, I will ignore matters of computational tractability and assume that the agent has a joint probability measure over the present state of the world, its past, and its future. I will continue to assume that time is linear and discrete. Any centred world in the algebra of propositions must then fix all relevant facts about the present state, its ancestors, and descendants.

We can therefore introduce the following piece of notation. For any centred world w and integer n, let w^{+n} be an otherwise identical world in which the centre is shifted n units of time into the future. How this is cashed out depends on the construction of centred worlds. For example, if centred worlds are triples of an uncentred world, an individual, and a time, then $(u, i, t)^{+n} = (u, i, t+n)$; if centred worlds are sentences, then the $+n$ operation prefixes the relevant sentence with 'in n units of time'. We can extend the operation to arbitrary propositions: for any set A of centred worlds,

$$A^{+n} = \{w^{+n} : w \in A\}.$$

The shifting operator $+n$ allows us to make explicit the agent's beliefs about how the world may change. For example, if the agent at t is 90% confident that it is going to rain in one unit of time given that it is raining now, then $\mathrm{Cr}_t(\mathit{rain}^{+1}/\mathit{rain}) = .9$. (This is what we would previously have represented as $\mathrm{Cr}^T(\mathit{rain}, \mathit{rain})$. The new representation also allows an agent's transition model to change over time.)

Now return to the two-step update described in (RN). The first step was to project the probabilities from t to $t+1$ by the transition model. The shifting operation makes this step easy to define: the new probability of any centred proposition A is simply the old probability of A^{+1}. For example, if Cr_t assigns probability .9 to rain^{+1} ('it is going to rain in 1 unit of time'), then the updated probability assigns probability .9 to *rain* ('it is raining now'). Let us denote the *shifted probability measure* that results from this step by Cr_t^{+1}. The concept obviously generalizes to larger (and smaller) intervals:

$$\mathrm{Cr}_t^{+n}(A) = \mathrm{Cr}_t(A^{+n}).$$

In the second step, the shifted probability measure is conditionalized on the evidence. If the evidence is captured by a (possibly centred) proposition E that is learnt with certainty, the whole update therefore looks as follows:

$$\mathrm{Cr}_{t+1}(A) = \mathrm{Cr}_t^{+1}(A/E) = \mathrm{Cr}_t(A^{+1}/E^{+1}). \tag{SC}$$

Since the second step is plain old conditionalization, generalizations to cases of uncertain and equivocal evidence are straightforward.

(SC) is the natural interpretation of (RN) if we assume that the objects of credence can be uncentred. In philosophy, (SC) has been defended in [Schulz 2010], [Meacham 2010], [Schwarz 2012], and [Schwarz 2015].[11] Schwarz also suggests that (SC) inherits many features of conditionalization if the probability space is extended by centred propositions.

11 Schulz does not isolate shifting as a separate step, and restricts (SC) to cases where the agent keeps track of time. (I will turn to this issue in the next section.) [Kim 2009] defends a similarly restricted norm with a similar form but where the relative shifts are replaced by shifts to an absolute time: $\mathrm{Cr}_{t+1}(A) = \mathrm{Cr}_t(A \textit{ at } i/E \textit{ at } i)$, provided the agent is certain at $t+1$ that the present time is i. Kim does not explain where the agent's new beliefs about the time come from.

For example, $\mathrm{Cr}_t^{+1}(\cdot/E)$ maximizes expected future accuracy, and agents are vulnerable to diachronic Dutch Books if and only if they violate (SC).

(SC) correctly deals with all our fruit basket examples. Recall that imaging accounts seemed to give the right result in the original case, assuming that the new information, that there are no bananas in the basket, is interpreted as information about an intervention: that somebody went ahead and removed any bananas from the basket. Since this is not equivalent to the simpler proposition that there are no bananas in the basket, we should include a corresponding event in the probability space. So the space should include events A (one apple, no banana), AA (two apples, no banana), AB (one apple, one banana), and RB (any bananas removed). To apply (SC), we also need A^{+1}, AA^{+1}, AB^{+1} and RB^{+1} so that we can model your beliefs at t about how the world might change. Let's assume that at time t you were neutral not only about whether there are two apples or an apple and a banana in the basket, but also about whether any bananas were going to be removed. Moreover, you regarded the two issues as independent, and you were certain that if the bananas won't be removed then all fruits will remain in the basket.[12] Then

$$\begin{aligned}
\mathrm{Cr}_t(AA^{+1} \wedge BR^{+1}) &= \mathrm{Cr}_t^{+1}(AA \wedge BR) = .25 \\
\mathrm{Cr}_t(AA^{+1} \wedge \neg BR^{+1}) &= \mathrm{Cr}_t^{+1}(AA \wedge \neg BR) = .25 \\
\mathrm{Cr}_t(A^{+1} \wedge BR^{+1}) &= \mathrm{Cr}_t^{+1}(A \wedge BR) = .25 \\
\mathrm{Cr}_t(AB^{+1} \wedge \neg BR^{+1}) &= \mathrm{Cr}_t^{+1}(AB \wedge \neg BR) = .25
\end{aligned}$$

Conditionalizing the shifted probability Cr_t^{+1} on BR leaves you with $\mathrm{Cr}_{t+1}(AA) = \mathrm{Cr}_{t+1}(A) = .5$. So (SC) still gets this case right. But it also applies, and gives sensible results, in cases where imaging accounts fall silent. For example, if BT is the proposition that a banana has been taken from the basket, $\mathrm{Cr}_t(AA^{+1} \wedge BT^{+1}) = 0$, and so $\mathrm{Cr}_{t+1}(A) = 1$.

So far, so good. But there are cases which (SC) seems to get wrong, notably if the agent loses track of time.

5 Losing track of time

If I have not just looked at a clock, I usually don't know the exact time. This suggests that I can't update my credences by (SC). To see why, imagine for the sake of vividness that the units of discrete time are minutes. Suppose at t an agent is certain that it is noon: $\mathrm{Cr}_t(\textit{12:00}) = 1$. Presumably, she can also be certain that in one minute it will be one minute past noon: $\mathrm{Cr}_t(\textit{12:01}^{+1}) = 1$. If the agent follows (SC), she will then be certain at one minute past noon that it is one minute past noon (assuming that she does

12 All these assumptions intuitively should affect the result of the update; it is an advantage of (SC) over (I) that they have to be made explicit.

not receive evidence with probability 0): $\mathrm{Cr}_{t+1}(\textit{12:01}) = \mathrm{Cr}_t(\textit{12:01}^{+1}) = 1$. In general, if an agent knows what time it is at any point and they update their beliefs in accordance with (SC), it seems that they will forever know the time.

Is this a problem? The classical norms of Bayesianism are norms for ideal agents with unlimited and perfectly reliable cognitive capacities. Real agents can't conditionalize complex probability measures in an instant; they can't compute the expected utility of all their options at every moment; they can't instantaneously see through all consequences of their beliefs; they can't retain every information they ever learned. All that does not undermine the relevant norms as constraints of ideal rationality. Nor does it make contemplating such norms pointless: ignoring cognitive limitations helps to simplify formal models (like ignoring friction and air resistance in physics); moreover, the ideal case can provide useful guidance when thinking about non-ideal cases.

So one might hold that my losing track of time is just a consequence of my cognitive limitations: cognitively ideal agents don't lose track of time.

The problem with this view is that it is false. Consider a time traveller who enters a time machine that he knows will take him either 100 or 200 years into the future. Upon arrival (without further evidence), the time traveller will be lost in time, even if she has unlimited cognitive resources. Similarly, consider an astronaut travelling to another star and back at a very fast but unknown speed. Arriving back on Earth, she can't know without further evidence how much time has passed on Earth, since that depends on the unknown velocity of her spaceship (according to Special Relativity). Or consider sleep—the poor man's version of time travel into the future. If you wake up from a surprising noise in the middle of the night, even unlimited cognitive resources won't help you to figure out the time. One might respond that cognitively ideal agents don't sleep, don't travel at unknown speed, and don't enter time machines without knowing the destination, but that is getting silly.

So we need to adjust (SC) to allow for agents who lose track of time. One way of doing that leads to the proposal of [Santorio 2011] and [Spohn Forthcoming]. On this account, (SC) is the right norm only if the agent is certain that one unit of time has passed since the earlier belief state. For the more general case, we assume that the agent can at least assign a probability to different hypotheses about how much time has passed. Suppose she gives 80% credence to the hypothesis that 1 unit of time has passed and 20% to the hypothesis that 2 units have passed. Instead of moving the probability mass of each centred world w to the corresponding world w^{+1}, the shifting step then divides the mass between w^{+1} and w^{+2}, with the former receiving 80% and the latter 20%. In general, let τ be a probability distribution over natural numbers representing the agent's belief about how many units of time has passed (still assuming that time is linear and discrete). To

shift the original credence function by τ, the probability of any world w is then given by

$$\mathrm{Cr}_t^\tau(w) = \sum_n \mathrm{Cr}_t(w^{+n})\tau(n). \tag{GI}$$

In the second step, the agent conditionalizes on her new evidence, like before:

$$\mathrm{Cr}_{t'}(A) = \mathrm{Cr}_t^\tau(A/E). \tag{SC$'$}$$

(I use '$\mathrm{Cr}_{t'}$' to denote the new credence function rather than 'Cr_{t+1}' because we no longer assume that the new credence is located 1 unit of time after Cr_t.)

To complete the proposal, we need to say where τ comes from. How does the agent arrive at a probability measure over how much time has passed since the earlier credence? The simplest answer, suggested by both Santorio and Spohn, is that τ comes from the agent's sensory evidence at t', delivered by a special "sense of time".[13]

Let us grant for the sake of discussion that all rational agents have a sense of time—although that strikes me as equally implausible as the assumption that rationality requires having eyes. Let us also not worry how the earlier time is represented by this sense—although I do worry about that.[14] The main problem with (SC$'$) is that it often gives verdicts that are clearly wrong.

Before I give an example, observe that (SC$'$) takes a step back towards the imaging approach. For if there is a proposition *that 1 unit of time has passed*, then Cr_t^{+1} is the image of Cr_t on that proposition: shifting turns into imaging. Indeed, (GI) is equivalent to generalized imaging as introduced in [Gärdenfors 1982] and [Lewis 1981] for cases in which there can be more than one "most similar" worlds. In contrast to the imaging accounts from section 3, however, (SC$'$) does not image the previous credence function on the total new evidence. The previous credence is imaged only on the information about how much time has passed. One might hope that this avoids the problems for the imaging account, because the information about how much time has passed—understood

13 [Moss 2012] also appeals to a sense of time for a similar purpose. Like (SC$'$), her proposal combines a shifting-like step with subsequent conditionalization; like Santorio and Spohn, she appeals to a sense of time for the first step. But on Moss's account, the first step does not take the form (GI). Instead, the agent completely discards her earlier self-locating beliefs and uses the sense of time (in a not fully explained manner) to locate herself in absolute time.

14 Spohn suggests that the sense of time delivers qualitative information, for example, that *2 minutes have passed since the red flag was waved* and *3 minutes have passed since the green flag was waved*. But what if the earlier time t is not distinguished by special features known with certainty to the agent? Moreover, how does the update process determine which of these qualitatively described times is the time of the earlier credence? According to Santorio and Moss, the earlier time is identified directly, *de re*. Here, too, I wonder how the update process knows which of the times given in that way is the time of the earlier credence.

as purely temporal information that doesn't even entail that the agent still exists—may be hoped to reveal nothing about what the world was like before.[15]

But a similar problem remains. Consider the following scenario.

> You're about to be put into an artificial coma for emergency surgery. If the surgery succeeds, you will wake up after a day. If it fails, you will wake up after 10 days. You know all this, and you rationally give credence 0.9 to the hypothesis that the surgery will fail. Your inner sense of time is not attuned to comas, so upon awakening it suggests to you that not much more than a day or two have passed. You have no further evidence at this point about whether the surgery succeeded.

Let's assume for concreteness that your sense of time assigns probability $1/2^n$ to the hypothesis that n days have passed; so $\tau(1) = 1/2$ and $\tau(10) = 1/1024$. Let S be the hypothesis that the surgery either will succeed or has succeeded, and let $D1$, $D2$, $D3$, ..., be the hypotheses that you were in the coma for 1, 2, 3, ..., days, respectively. At t, before you were put into the coma, your credence in S was 0.1. Shifting by (GI) leads to the following probabilities:

$$\begin{gathered}
\mathrm{Cr}_t^\tau(S \wedge D1) = 0.1 \times 1/2 = 0.05 \\
\mathrm{Cr}_t^\tau(\neg S \wedge D1) = 0.9 \times 1/2 = 0.45 \\
\cdots \\
\mathrm{Cr}_t^\tau(S \wedge D10) = 0.1 \times 1/1024 \approx 0.0001 \\
\mathrm{Cr}_t^\tau(\neg S \wedge D10) = 0.9 \times 1/1024 \approx 0.0009. \\
\cdots
\end{gathered}$$

Notice that the shifted probabilities assign significant mass to possibilities like $\neg S \wedge D1$ in which you are not awake. (If we replace the time in the coma with genuine time travel, you wouldn't even exist there.) These possibilities must be ruled out in the next step of the update, where you conditionalize. Your post-awakening evidence plausibly reveals to you that you are awake and that you have just woken up from the coma. Thus it concentrates all probability on $S \wedge D1$ and $\neg S \wedge D10$. As a result, you will become highly confident that the surgery was a success:

$$\begin{aligned}
\mathrm{Cr}_{t'}(S \wedge D1) &\approx 0.9827 \\
\mathrm{Cr}_{t'}(\neg S \wedge D10) &\approx 0.0173.
\end{aligned}$$

15 Arguably, the information that so-and-so much time has passed still entails that the world hasn't come to an end in the meantime, and thus reveals that the previous state of the world was not a terminal state. Terminal world states also pose problems for (SC): what is w^{+1} if w is terminal? I will ignore these problems and assume that the agents in question give negligible credence to the hypothesis that the world is just about to end.

This is the wrong result. Knowing that your sense of time is not to be trusted, your credence in the surgery having been a success should remain close to the previous level, at around 0.1.

The coma example is not an isolated special case. Our inner sense of time is not perfectly calibrated to the actual passage of time, and we know that its reliability, and the direction in which it errs, depends on the circumstances: during some activities, time seems to fly by, during others it almost comes to a standstill. We cannot blindly trust our sense of time.

What if we don't identify τ in (SC′) with the output of an agent's sense of time, but with the agent's all-things-considered credence about how much time has passed, based on her sense of time together with any other relevant information she may have? (SC′) then collapses into contradiction. To illustrate, return to the coma case and suppose at t' your all-things-considered credence in $D1$ and $D10$ is 0.1 and 0.9 respectively (as seems reasonable). (GI) then moves probability $0.9 \times 0.9 = 0.81$ to $\neg S \wedge D10$ worlds, and $0.1 \times 0.1 = 0.01$ to $S \wedge D1$ worlds. Conditionalizing on the information that you've woken up excludes all other possibilities, so $\mathrm{Cr}_{t'}(D10) \approx 0.988$ and $\mathrm{Cr}_{t'}(D1) \approx 0.012$—contradicting the assumption about your all-things-considered credence in $D1$ and $D10$.[16]

[Schulz 2010] suggests a different generalization of (SC) that gets around the problem. Let τ represent the agent's all-things-considered credence (at t') about how much time has passed. Then Schulz suggests to replace (SC) by (SC*):

$$\mathrm{Cr}_{t'}(A) = \sum_n \tau(n)\mathrm{Cr}_t(A^{+n}/E^{+n}). \qquad \text{(SC*)}$$

In the coma example, $\mathrm{Cr}_t(S^{+1}/E^{+1})$ is plausibly 1 and $\mathrm{Cr}_t(S^{+10}/E^{+10})$ is 0, for you are sure that the surgery succeeds iff you awaken after one day. With $\tau(1) = 0.1$ and $\tau(10) = 0.9$, we therefore get the desired result:

$$\mathrm{Cr}_{t'}(S) = 0.1 \times 1 + 0.9 \times 0 = 0.1.$$

In contrast to (SC) and (SC′), (SC*) can no longer be divided into a shifting step and a conditionalization step.[17]

16 The only way to get $\mathrm{Cr}_{t'}(D10) = 0.9$ and $\mathrm{Cr}_{t'}(D1) = 0.1$ out of (SC′) is to assume that τ assigns equal probability to $D10$ and $D1$. In general, (SC′) often gives plausible verdicts if τ is uniform over all possible time shifts. The shifting step then makes the agent completely lost in time; to find her new location in time, the agent must draw on her (remaining) evidence. With uniform τ, (SC′) belongs to the family of proposals discussed in section 2 on which diachronic norms only pertain to uncentred beliefs.

17 [Kim 2009] proposes a similar generalization to (SC) but with absolute time shifts: $\mathrm{Cr}_{t'}(A) = \sum_i \mathrm{Cr}_{t'}(now = i)\mathrm{Cr}_t(A \text{ at } i/E \text{ at } i)$, where i ranges over points of time. (Notice that this only makes use of the agent's uncentred opinions at t.) The following counterexample to (SC*) is also a counterexample to Kim's proposal.

Schulz's proposal escapes the coma problem, but it gives wrong results in other cases, such as the following variant of the Sleeping Beauty problem (not involving any threat of memory erasure).

> After Beauty goes to sleep on Sunday, a fair coin is tossed. If it lands heads, Beauty will be made to sleep through until Tuesday morning when she will be awakened by the sound of a bell; if the coin lands tails, she will be awakened on Monday by the cry of a rooster and on Tuesday by the sound of a bell. Beauty knows all this when she goes to sleep. Then the coin is tossed and lands heads.

It should be uncontroversial that when Beauty awakens to the sound of a bell, she ought to become confident that it is Tuesday and the coin landed heads. So, using days as temporal units, $\tau(1) = 0$ and $\tau(2) = 1$; by (SC*), it follows that $Cr_{Tue}(\mathit{Heads}) = 1 \times \mathrm{Cr}_{Sun}(\mathit{Heads}^{+2}/E^{+2})$. But Beauty already knew on Sunday that she would be awakened by a bell in two days time. So if her relevant sensory evidence E is the sound of the bell (or her awakening by that sound), then $\mathrm{Cr}_{Sun}(\mathit{Heads}^{+2}/E^{+2}) = 1/2$, and so (SC*) falsely entails that upon awakening, Beauty should give equal credence to heads and tails.[18]

Another obvious drawback of (SC*) is that it takes the agent's all-things-considered credence about how much time has passed as given. How is the agent supposed to arrive at these beliefs? Not by (SC*). But beliefs about how much time have passed are highly constrained by the agent's previous belief state and her sensory evidence. As it stands, (SC*) is therefore at best part of the full story about ideal diachronic rationality—even setting aside that it gets some cases wrong and relies on a questionable ability to pick out the previous time.[19]

Fortunately, we can say more. To do so, we need to have a closer look at the space of centred propositions.

6 Shifting in personal time

Return to the simple shifting rule from section 4:

$$\mathrm{Cr}_{t+1}(A) = \mathrm{Cr}_t^{+1}(A/E) = \mathrm{Cr}_t(A^{+1}/E^{+1}). \tag{SC}$$

So far, we have interpreted the rule as specifying how an agent's credences should evolve from any time t to the next time $t+1$. In effect, this assumes that the algorithm for

18 To be fair, Schulz does not say whether E is to be understood as sensory evidence or as a more comprehensive kind of evidence that includes the information that Beauty was not awakened by a rooster the day before (which, as I hereby stipulate, is not part of her sensory evidence).

19 [Titelbaum 2016] points out the analogous gap in [Kim 2009] and falsely claims that it affects all shifting accounts.

updating the agent's beliefs is executed at every unit of time. But that is hardly a norm of rationality. Indeed, if a time traveller travels (instantaneously) from t to some point t' in the distant past or future, her beliefs upon arrival at t' should be updated from her beliefs at t, not from her beliefs at $t'-1$, when she didn't even exist.

So we should not make any fixed assumptions about how the post-update belief state and the pre-update state are related in "external time", as measured by atomic clocks. We should not assume that the post-update state is located precisely one unit of time after the pre-update state. On the other hand, we also don't need to look for norms relating arbitrary belief states at arbitrary times. ([Santorio 2011] and [Spohn Forthcoming] defend (SC′) as a general norm of this kind.) That would complicate our task because sensory evidence is not cumulative: your sensory evidence at t_1 is generally not part of your sensory evidence at t_2. Thus to say how your t_2 state should be related to your t_0 state, we need to take into account the intermediate evidence at t_1 and its impact on your rational credence. In any case, it is plausible that if you rationally update your credences from t_0 to t_1 and then from t_1 to t_2, then your t_2 credences stand in the right epistemic relation to your t_0 credences. Thus we may still focus on norms for *successive* belief states. It's just that we shouldn't assume that these states are separated by a particular distance in external time. The pre-update state is given not as the state from 1 unit of time in the past, but simply as *the state that is updated*.

To simplify the technical details, I will still assume that we're dealing with agents who update their belief state instantaneously and in discrete steps. (The continuous generalizations are mathematically routine.) I will also assume that an update always takes precisely one belief state as input and produces precisely one new belief state, thus ignoring cases of personal fission and fusion. So we can always speak of 'the previous belief state', 'the next belief state', 'the belief state before the previous belief state', etc. This gives us a kind of *personal time* (compare [Lewis 1976a]) in which, by definition, one unit of time passes with every update step.

We can interpret the shifting rule (SC) as a rule for updating in personal time, specifying how an ideal update process should transform an agent's belief state, thereby creating its successor state, no matter how the two states are related in external time. This requires adjusting the interpretation of the shifting operation $+1$. Where previously w^{+1} was a world just like w but with the centre shifted one unit of external time into the future, the shift now goes by personal time. How that is cashed out again depends on how w is construed. If w is a triple of an uncentred world, an individual, and a time, then w^{+1} may be construed as the triple of the same individual in the same world, but with the time shifted to the next point in the epistemic history of that individual at that world; that is, the point at which the individual next updates her belief state. If w is a sentence, then w^{+1} is that sentence prefixed by something like 'at the next belief update'.[20]

20 This is (roughly) how [Meacham 2010] and [Schwarz 2012] interpret (SC).

With these changes, (SC) gives the right verdicts even in time travel cases and situations where agents lose track of time, so the pressure to explore more general rules like (SC′) or (SC*) disappears.

Here is how the re-interpreted version of (SC) applies in the coma example. The shifting step moves the probability of every pre-coma world to the next post-coma world. For worlds in which the surgery is a success, the shift covers 1 day; for worlds in which the surgery fails, it covers 10 days. (By contrast, (SC′) indiscriminately shifted all worlds by the same amount.) The shifted credence thus still assigns probability 0.1 to the surgery having succeeded, and since your new evidence is neutral on the success of the surgery, this is your final credence.

In the variant of the Sleeping Beauty scenario, the shifted Sunday credence is evenly divided between *Heads & Tuesday & Bell* worlds and *Tails & Monday & Rooster* worlds; subsequent conditionalizing on *Bell* makes Beauty certain that it is Tuesday and the coin landed heads.

(SC) does not deny that we have a sense of time, but it also doesn't make it a requirement of rationality. If an agent has such a sense, it enters the update process like any other sense, in the conditionalization step.

One might worry that (SC), on its present interpretation, requires the agent to have beliefs about her successor belief states: about whether A will be true at the time of the next belief state given that E is true then. Are such beliefs really well-defined? Must rational agents really have them?

I have already explained how talk about successor belief states is to be understood. The assumption that makes the notion well-defined is that agents update their beliefs in discrete steps, and that we set aside cases of fission and fusion (and ignore terminal states). To make the agent's credences over propositions like A^{+1} and E^{+1} well-defined, we have to further assume that these assumptions hold throughout the agent's doxastic space. For many concrete applications, this is not a serious constraint, but generalizations are needed (and can be provided) for cases in which some of the assumptions must be dropped.

The second worry I take more seriously. Must rational agents really have beliefs about their successor states? Couldn't a perfectly rational agent have no beliefs at all about her own psychology and its dynamics? A simple response is that if the optimal update require an agent's doxastic space to include certain propositions, and these propositions are meaningful, then ideal rationality does indeed require an agent's credences to be defined over such propositions. This kind of lesson is not new. It is well-known that conditionalization, for example, yields wrong results if an agent's credences are defined over too narrow a space.

Moreover, we saw in section 3 that a plausible update account must take into account not only the agent's beliefs about the present state of the world, but also her views

about how that state is likely to change over time. In this context, the epistemically more important kind of time is arguably not external time, but the agent's personal time. In simple environments, an agent's subjective probability space may therefore ignore external time: the probabilities may be defined simply over world states ordered by personal time. And of course a unit of personal time does not have to be conceptualized as 'the time when I next update my beliefs'.

7 Cumulative evidence

To conclude, I want to briefly return to the views from section 2 that do not impose diachronic norms on self-locating beliefs—in some cases by rejecting the very idea that the objects of credence may be centred. These views are often motivated by the observation that conditionalization does not apply to centred propositions, and a suspicion that no suitable replacement norm can be found. Stalnaker, for example, complains that "centred-worlds models [...] provided no resources for representing the relations between informational states across time and across persons, and so no resources for clarifying the dynamics of knowledge and belief" [Stalnaker 2008: 64]. As we have seen, this suspicion is unfounded, at least for rational degree of belief: the centred-worlds model allows us to specify precisely how ideal credence evolves over time.

As I mentioned in section (2), many authors accept the centred-worlds model of rational credence, but suggest that diachronic norms should be replaced by synchronic norms. (Again, this is often motivated by the supposed difficulty of finding diachronic norms.) On the simplest proposal, an agent's credence function at any time should equal some ur-prior P conditional on the agent's evidence at that time, irrespective of the agent's previous beliefs. I have argued that these views require an unexplained non-sensory conception of evidence and thus don't provide a full answer if we want to know how an agent's credences should evolve as time goes by and new information arrives from her senses. Moreover, I suggested that if the accounts are to be plausible, they cannot avoid the problem of updating centred information. Returning to our original example, if you fall asleep while listening to the rain, then your non-sensory evidence upon awakening should include the information that it was raining earlier; your earlier sensory evidence that it *is* raining turns into your later non-sensory evidence that is *was* raining. What are the general rules for this update process on an agent's evidence?

The discussion of the previous sections suggests a natural answer: if E is your evidence at some point, then your non-sensory evidence at the next point in your personal time should be E^{-1} conjoined with whatever new information you acquire through your senses. E^{-1} is the centred proposition that E was the case one unit of personal time in the past.

Assume an agent receives sensory evidence $E_1, E_2, E_3, \ldots, E_n$ at successive points in her personal time. Define the agent's *cumulative evidence* at the last point as

$E_n \wedge E_{n-1}^{-1} \wedge E_{n-2}^{-2} \wedge \ldots \wedge E_1^{-(n-1)}$. In classical Bayesianism, where centred propositions are ignored, the result of successively updating on E_1, E_2, E_3, etc., is always identical to conditionalizing the initial credence function on the conjunction of E_1, E_2, E_3, etc. One might similarly conjecture that successive application of (SC) to an agent's sensory evidence leads to the same result as conditionalizing the initial credence function on the agent's final cumulative evidence.

The conjecture turns out to hold only under non-trivial background assumptions. In particular, as I show in the appendix, we need the following *stationarity assumption* (a special case of the principle of self-locating indifference defended in [Elga 2004]):

> Whenever a centred world w' is a successor of a world w', then the initial credence Cr_0 assigns equal probability to w and w'.

If one accepts stationarity and identifies an agent's non-sensory evidence with their cumulative sensory evidence, then successive updating on sensory evidence in accordance with (SC) and ur-prior conditionalization (with the initial credence Cr_0 as ur-prior) yield the same result. One might prefer the ur-prior formulation because it more easily generalizes to cases where agents lose information. My own view is that it does not, and that rationality sometimes requires violating stationarity, but that is a story for another occasion.

Appendix

I will show that if stationarity holds (as well as a minor further assumption that I will introduce in a moment), then successive updating on sensory evidence in accordance with (SC) yields the same result as conditionalizing the agent's initial credence Cr_0 on her cumulative evidence.

Let $E_1, \ldots, E_n$ be the sensory evidence the agent receives at personal times $1, \ldots, n$, respectively. With a little algebra, it is easy to show that if the agent follows (SC), then for any proposition A,

$$\mathrm{Cr}_n(A) = \mathrm{Cr}_0(A^{+n}/E_1^{+1} \wedge \ldots \wedge E_n^{+n}). \tag{1}$$

Since the agent's cumulative evidence at point n is $E_1^{-(n-1)} \wedge \ldots \wedge E_{n-1}^{-1} \wedge E_n$, what we have to show is (2).

$$\mathrm{Cr}_0(A^{+n}/E_1^{+1} \wedge \ldots \wedge E_n^{+n}) = \mathrm{Cr}_0(A/E_1^{-(n-1)} \wedge \ldots \wedge E_{n-1}^{-1} \wedge E_n). \tag{2}$$

The further assumption we need besides stationarity is that E_1 is incompatible with the present state being an agent's initial state, which can be captured by (3), where $\top$ is the tautology:

$$Cr_0(E_1 \supset \top^{-1}) = 1. \tag{3}$$

To see why (3) is needed, suppose Cr_0 assigns

- probability .4 to an initial $A \wedge E_1$ world,
- probability .4 to its successor, a terminal $\neg A \wedge \neg E_1$ world,
- probability .1 to an initial $A \wedge \neg E_1$ world, and
- probability .1 to its successor, a terminal $\neg A \wedge E_1$ world.

Then $\mathrm{Cr}_0(A^{+1}/E_1^{+1}) = 0$ but $\mathrm{Cr}_0(A/E) = .8$, which violates (2). (3) could perhaps be motivated by the fact that an agent's total new evidence generally includes higher-order evidence about the agent's beliefs. Alternatively, instead of (3) we could assume that even an agent's initial credence Cr_0 is given by conditionalizing a merely hypothetical ur-prior P on some initial evidence E_0, which would play the role of $\top$ in what follows.

To prove (2), observe that any centred world that satisfies $A_0 \wedge A_1^{+1} \wedge \ldots \wedge A_n^{+n}$, for any propositions $A_0, A_1, \ldots, A_n$, is succeeded by a world that is succeeded by a world ... (n times) ... that satisfies $A_0^{-n} \wedge A_1^{-(n-1)} \wedge \ldots \wedge A_n$; conversely, any world that satisfies $A_0^{-n} \wedge A_1^{-(n-1)} \wedge \ldots \wedge A_n$ is n-preceded by a world that satisfies $A_0^{-n} \wedge A_1^{-(n-1)} \wedge \ldots \wedge A_n$; by linearity (no fission and fusion), this mapping from worlds to worlds is one-one; by stationarity, the worlds it pairs always have equal probability. Thus for any propositions A_0, $A_1, \ldots, A_n$,

$$Cr_0(A_0 \wedge A_1^{+1} \wedge \ldots \wedge A_n^{+n}) = Cr_0(A_0^{-n} \wedge A_1^{-(n-1)} \wedge \ldots \wedge A_n). \tag{4}$$

Two instances of this equality are (5) and (6).

$$\mathrm{Cr}_0(\top \wedge E_1^{+1} \wedge \ldots \wedge E_n^{+n}) = \mathrm{Cr}_0(\top^{-n} \wedge E_1^{-(n-1)} \wedge \ldots \wedge E_n) \tag{5}$$

$$\mathrm{Cr}_0(\top \wedge E_1^{+1} \wedge \ldots \wedge E_n^{+n} \wedge A^{+n}) = \mathrm{Cr}_0(\top^{-n} \wedge E_1^{-(n-1)} \wedge \ldots \wedge E_n \wedge A). \tag{6}$$

By (3), $Cr_0(E_1 \leftrightarrow (\top^{-1} \wedge E_1)) = 1$, so we can remove $\top$ from the conjunctions in (5) and (6). By the ratio formula for conditional probability, (5) and (6) then straightforwardly entail (2).

References

Carlos E. Alchourrón, Peter Gärdenfors and David Makinson [1985]: "On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision". *Journal of Symbolic Logic*, (50): 510–530

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

Nick Bostrom et al. [2002]: *Anthropic bias: Observation selection effects in science and philosophy*. New York: Routledge

Craig Boutilier [1998]: "A unified model of qualitative belief change: a dynamical systems perspective". *Artificial Intelligence*, 98: 281–316

Darren Bradley [2013]: "Dynamic Beliefs and the Passage of Time". In A. Capone and N. Feit (Eds.) *Attitudes De Se,* Chicago: University of Chicago Press

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

David Chalmers [2011]: "Frege's puzzle and the objects of credence". *Mind*, 120(479): 587–635

David Christensen [2000]: "Diachronic coherence versus epistemic impartiality". *The Philosophical Review*, 109(3): 349–371

Adam Elga [2004]: "Defeating Dr. Evil with Self-Locating Belief". *Philosophy and Phenomenological Research*, 69: 383–396

Peter Gärdenfors [1982]: "Imaging and Conditionalization". *Journal of Philosophy*, 79: 747–760

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

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

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

Terry Horgan [2008]: "Synchronic Bayesian Updating and the Sleeping Beauty Problem: Reply to Pust". *Synthese*, 160: 155–159

Richard Jeffrey [1965]: *The Logic of Decision.* New York: McGraw-Hill

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

Gabriele Kern-Isberner [2001]: "Revising and Updating Probabilistic Beliefs". In *Frontiers in Belief Revision,* Springer, 393–408

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

Jérôme Lang [2007]: "Belief Update Revisited." In *IJCAI,* vol 7. 6–12

Hannes Leitgeb [2016]: "Imaging all the people". *Episteme*: 1–17

Hannes Leitgeb and Richard Pettigrew [2010]: "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy". *Philosophy of Science*, 77: 236–272

Isaac Levi [1980]: *The Enterprise of Knowledge.* Cambridge, MA: MIT Press

David Lewis [1976a]: "The Paradoxes of Time Travel". *American Philosophical Quarterly*, 13: 145–152

— [1976b]: "Probabilities of Conditionals and Conditional Probabilities". *The Philosophical Review*, 85: 297–315

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

— [1981]: "Causal Decision Theory". *Australasian Journal of Philosophy*, 59: 5–30

— [1999]: "Why Conditionalize?" Cambridge: Cambridge University Press, 403–407

Christopher Meacham [2008]: "Sleeping Beauty and the Dynamics of De Se Beliefs". *Philosophical Studies*, 138: 245–269

— [2010]: "Unravelling the Tangled Web: Continuity, Internalism, Non-Uniqueness and Self-Locating Beliefs". In Tamar Szabo Gendler and John Hawthorne (Eds.) *Oxford Studies in Epistemology, Volume 3,* Oxford University Press, 86–125

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

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

— [2014]: "Time-slice epistemology and action under indeterminacy". *Oxford Studies in Epistemology*, 5

John Perry [1977]: "Frege on Demonstratives". *Philosophical Review*, 86: 474–497

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

Alvin Plantinga [1974]: *The Nature of Necessity.* Oxford: Oxford University Press

John L. Pollock [1984]: *The foundations of philosophical semantics.* Princeton: Princeton University Press

Stuart J. Russell and Peter Norvig [2010]: *Artificial Intelligence: A Modern Approach*. Cambridge (MA): MIT Press, 3rd edition

Nathan Salmon [1986]: *Frege's Puzzle*. Cambridge (Mass.): MIT Press

Paolo Santorio [2011]: "Cognitive Relocation". Unpublished Manuscript

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

Wolfgang Schwarz [2012]: "Changing Minds in a Changing World". *Philosophical Studies*, 159: 219–239

— [2015]: "Belief update across fission". *British Journal for the Philosophy of Science*, 66: 659–682

Charles T Sebens and Sean M Carroll [2017]: "Self-locating uncertainty and the origin of probability in Everettian quantum mechanics". *The British Journal for the Philosophy of Science*

Brian Skyrms [1987]: "Dynamic coherence and probability kinematics". *Philosophy of Science*, 54(1): 1–20

Wolfgang Spohn [Forthcoming]: "The Epistemology and Auto-Epistemology of Temporal Self-Location and Forgetfulness". *Ergo*

Robert Stalnaker [1981]: "Indexical Belief". *Synthese*, 49: 129–151

— [2008]: *Our Knowledge of the Internal World*. Oxford: Oxford University Press

— [2014]: *Context*. Oxford: Oxford University Press

— [2016]: "Modeling a Perspective on the World". In Manuel García-Carpintero and Stephan Torre (Eds.) *About Oneself: De Se Thought and Communication,* Oxford: Oxford University Press, 121–139

Paul Teller [1973]: "Conditionalization and observation". *Synthese*, 26(2): 218–258

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

— [2013]: *Quitting Certainties*. Oxford: Oxford University Press

— [2016]: "Self-Locating Credences". In A. Hajek and C. Hitchcock (Eds.) *The Oxford Handbook of Probability and Philosophy,* Oxford: Oxford University Press

Bernard Walliser and Denis Zwirn [2002]: "Can Bayes' Rule be Justified by Cognitive Rationality Principles?" *Theory and Decision*, 53(2): 95–135

Clas Weber [2013]: "Centered communication". *Philosophical Studies*, 166(1): 205–223

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