Evidentialism and Conservatism in Bayesian Epistemology*

Wolfgang Schwarz
Draft, 30 January 2017

What is the connection between evidential support and rational degree of belief? It is often taken for granted that that the degree of belief we assign to a hypothesis ought to match the extent to which the hypothesis is supported by our evidence. I show that this "evidentialist" assumption sometimes clashes with the equally intuitive "conservative" assumption that we should not change our beliefs unless there is a reason to do so. I also suggest that in cases where the evidentialist and the conservative norm come apart, there are reasons to side with conservatism.

1 Introduction

"A wise man", said Hume, "proportions his beliefs to the evidence" [Hume 1777/1993: sec.X, part I]. Using the Bayesian notion of graded belief, we might paraphrase this *evidentialist* principle as follows: the degree of belief a rational agent assigns to a hypothesis matches the extent to which the hypothesis is supported by the agent's evidence. Here we assume that one can measure the extent to which a hypothesis is supported by an agent's evidence. Not everyone would accept that; but if the assumption is granted, the evidentialist principle appears to state an almost trivial connection between rational degree of belief and evidential support.

Another platitude about rational belief is what I will call the principle of *(doxastic) conservatism*. Informally, the principle says that one should not change one's beliefs unless there is a reason to do so. More specifically, suppose at some point an agent learns some proposition E and nothing else; if A is any other proposition which the agent previously and rationally regarded as independent of E, then we may take the principle of conservatism to say that the event of learning E should not change the agent's degree of belief in A.[1]

The main aim of the present paper is to show that these two platitudes – the principle of evidentialism and the principle of conservatism – can pull in opposite directions. At

* Ancestors of this paper were presented at the 2012 AAP in Wollongong and at the University of Saarbrücken in 2013. I thank the audiences for comments and discussion.

1 As stated, the principle of conservatism falsely presupposes new information is the only rational reason to change one's degrees of belief; see section 5.

least one of them must be given up. I will suggest that we give up the principle of evidentialism: a wise man (or woman) does not always proportion his (or her) beliefs to the evidence.

That the two principles can come apart is not surprising if we think about the constraints they impose on rational degrees of belief. The principle of evidentialism implies that rational degrees of belief are fully determined by the agent's present evidence: if two agents have the same evidence, they ought to have the same degrees of belief. The principle of conservatism, on the other hand, implies that rational degrees of belief are constrained not only by present evidence but also by previous beliefs. We may therefore expect the two principles to come apart in cases where the previous beliefs are not recoverable from present evidence. A well-known scenario of this kind is the Sleeping Beauty problem. Indeed, in section 7 I argue that the disagreement between halfing and thirding has its roots in the disagreement between conservatism and evidentialism: conservatism supports halfing, evidentialism thirding.

However, conservatism and evidentialism don't merely come apart in the treatment of memory loss. In section 6 I discuss a situation in which an agent faces the possibility of personal fission. Cases like this have played an important role in the debate over Everettian Quantum Mechanics. We will see that here, too, evidentialism and conservatism give interestingly different recommendations.

In section 5 I argue that the two positions further disagree over Adam Elga's [2004] Dr. Evil scenario: from a conservative perspective, Dr. Evil's belief that he is Dr. Evil should not be swayed by learning he has a duplicate with the exact same evidence.

In section 4 I consider what happens if we relax the evidentialist *Uniqueness thesis*, that "given one's total evidence, there is a unique rational doxastic attitude one can take to any proposition" [White 2005: 455]. If we move to a *permissive* form of evidentialism on which an agent's evidence sometimes allows for a whole range of doxastic states, further opportunity arises for a disagreement with conservatism.

That diachronic and evidential norms of rationality can clash has been noted before (e.g. in [Arntzenius 2003], [Moss 2015b], [Hedden 2015b], [Hedden 2015a]), but the scale of the disagreement has not been properly appreciated. Moreover, those who noted the clash have generally inferred that the relevant diachronic norms should be rejected. In section 8 I look into this inference and argue that it rests on a misunderstanding of diachronic norms.

The observations of the present paper have ramifications for other debates, although I do not have the space to explore these here. For example, if I am right there can be disagreement between perfectly rational agents with the very same priors and the very same evidence – which contradicts a common assumption in the literature on disagreement. For another example, the dominant theory of scientific confirmation, Subjective Bayesian Confirmation Theory, assumes that the degree to which a hypothesis is supported by

evidence can be identified with the rational degree of belief agents would assign to the hypothesis on the basis of this evidence. Again, the observations of the present paper imply that this is not generally true.

The positions I call evidentialism and conservatism are obviously related to their namesakes in traditional (non-Bayesian) epistemology. Here, Richard Feldman and Earl Conee [1985: 15] influentially characterized evidentialism as the view that "[d]oxastic attitude D towards proposition p is epistemically justified for S at t if and only if having D towards p fits the evidence S has at t"; Hamid Vahid [2004: 97] describes conservatism as the view that "it would be unreasonable to change one's beliefs in the absence of any good reasons". However, the connection between these views on the reasonableness or justification of all-or-nothing belief and the Bayesian doctrines that are my focus is not entirely straightforward, and I will leave the ramifications for these traditional views unexplored.

In the next two sections, I will begin by setting the stage. I will review the basic tenets of Bayesian epistemology and explain how evidentialism and conservatism may be expressed in that framework.

2 Conservatism and conditionalization

Bayesian epistemology replaces the threefold distinction between belief, disbelief and suspension of judgement with a more fine-grained scale of degrees of belief or credences. The idea has proved fruitful not only in epistemology, but also in neighbouring disciplines such as confirmation theory ([Earman 1992]), statistics ([Howson and Urbach 1993]), decision theory ([Joyce 1999]), artificial intelligence ([Russell and Norvig 2004]), and cognitive science ([Oaksford and Chater 2007]).

Two norms on rational credence form the core of classical Bayesian epistemology. The first is *probabilistic coherence*, which holds that credences should obey the laws of the probability calculus. The second, *conditionalization*, is a diachronic norm on the evolution of rational credence. In its simplest form, the norm says that if at some time t an agent becomes certain of some proposition E_t (and nothing else), then her credence in any proposition A should equal her previous credence in A conditional on E_t:

$$Cr_t(A) = Cr_{t-1}(A/E_t).$$

Here 'Cr_{t-1}' denotes the agent's credence function just before she learned E. The conditional probability on the right-hand side is often computed via Bayes' Theorem, which is why conditionalization is also known as *Bayes' Rule*.

Conditionalization only applies in cases where the direct impact of a learning event (typically, a sensory experience) on an agent's degrees of belief is that some proposition E_t becomes certain. Charitably understood, classical Bayesianism does not say that learning

events always work like that, or that there are no other reasons to change one's degrees of belief. More complicated rules have been proposed for situations where learning events impose weaker constraints on the new probabilities (see [Jeffrey 1965], [Skyrms 1980], [Bradley 2005]), for the dynamics of belief in practical deliberation ([Skyrms 1990]), and for updating an agent's beliefs about her own location in the world, a topic to which we will return in section 5. For the moment, let us focus on simple cases where an agent's experiences render some proposition certain and the agent has no other reasons to change her beliefs.

These are also the conditions under which the principle of conservatism, as formulated in the previous section, applies. Recall: when an agent learns some proposition E and nothing else, then her degree of belief in any other proposition which she previously and rationally regarded as independent of E should remain unchanged. The principle is entailed by the norm of conditionalization: if an agent conditionalizes on E_t and her previous credence in A was equal to her conditional credence in A given E_t (i.e., she treated the two propositions as independent), then her credence in A won't change. Indeed, a well-known argument for conditionalization is precisely that it satisfies various "minimal revision" constraints such as our principle of conservatism (see e.g. [Teller 1973], [Williams 1980]).

So classical Bayesianism is conservative. Given the tight connection between conditionalization and conservatism, various arguments for conditionalization can be seen as indirect arguments for the principle of conservatism. For example, David Lewis (reported in [Teller 1973]) and Brian Skyrms [1987] showed that an agent is vulnerable to diachronic Dutch Books if and only if her beliefs do not change by conditionalization. That is, if (and only if) an agent does not revise her beliefs by conditionalization, one can construct a series of bets – some offered before the learning event, some after – all of which will appear favourable to the agent but which in combination amount to a sure loss. Since conditionalization implies the principle of conservatism, violating the principle of conservatism implies vulnerability to diachronic Dutch Books. So if we regard vulnerability to Dutch Books as a sign of irrationality – a controversial premise, to be sure – we should accept the principle of conservatism as a rational norm.

Most Bayesians accept further epistemic norms besides probabilistic coherence and conditionalization, such as David Lewis's [1980] "Principal Principle". These are often expressed as constraints on *initial* credence functions, before any contingent information has been learnt. The point of such norms is not to make prescriptions for real agents who don't have any evidence (newborns? embryos?). Rather, the point is to distinguish two kinds of deviation from ideal rationality: inadequate accommodation of new information and inadequate prior beliefs. Observe that conditionalizing on E_1 and then on E_2 has the same effect as conditionalizing in one step on the conjunction of E_1 and E_2: if $Cr_2 = Cr_1(\cdot/E_2)$ and $Cr_1 = Cr_0(\cdot/E_1)$, then $Cr_2 = Cr_0(\cdot/E_1 \wedge E_2)$. So if an agent

always revises her beliefs by conditionalization, then her credence function at any time t equals her initial credence function Cr_0 conditional on the conjunction of all her evidence up to t:

$$Cr_t(A) = Cr_0(A/E_1 \wedge \ldots \wedge E_t).$$

In the presence of conditionalization, all other norms on rational credence therefore map onto constraints on the initial credence function Cr_0.

Bayesians disagree on how tightly the norms of rationality constrain initial credence (see e.g. [Meacham 2014]). Opinions range from "radical subjectivism", on which any coherent credence function is permissible, to what might be called "radical objectivism", on which there is only one rational initial credence function. The latter view will be useful to set up the comparison between conservative Bayesianism and its evidentialist rival.

3 Evidentialism

Evidentialism says that the degree of belief we assign to a proposition should match the extent to which that proposition is supported by our evidence. This assumes that one can measure the extent to which propositions are supported by evidence, and that the relevant measure can be shoehorned into a (conditional) probability measure. Probabilistic measures of evidential support are indeed popular in confirmation theory, but here the probabilities are often identified with subjective degrees of belief. That is, the degree to which some evidence ("absolutely") confirms some hypothesis, relative to some agent, is identified with the agent's degree of belief in the hypothesis given the evidence. But clearly the evidentialist principle is meant to say more than that our degree of belief in any proposition should equal our degree of belief in that proposition conditional on our evidence. The principle therefore seems to require a form of "objective Bayesian confirmation theory", where the confirmation measure is not identified with anyone's actual degrees of belief (see e.g. [Hawthorne 2005], [Maher 2010], [Williamson 2011]). So let's assume that there is an objective conditional probability measure Conf that captures the extent to which any proposition A is supported by any evidence proposition E. Let's also assume for now that Conf is fully precise and determinate – we will relax this assumption in section 4. The principle of evidentialism can then be expressed as the claim that if an agent's total evidence at time t is E_t, then her credence $Cr_t(A)$ in any

proposition A should equal $\mathrm{Conf}(A/E_t)$:[2]

$$Cr_t(A) = \mathrm{Conf}(A/E_t).$$

Qualms about objective standards for evidential support are one reason to question evidentialism. I want to focus on a different reason. I will argue that even if we grant the idea of objective evidential support, evidentialism should be rejected because it clashes with conservatism. For the present paper, I therefore want to largely set aside qualms about objective evidential support.

On any account there must be *some* connection between evidential support (if there is such a thing) and rational belief. In the classical Bayesian picture, the connection is naturally understood in terms of constraints on initial credence. Specifically, if there is a unique and precise objective confirmation measure Conf, then one's rational initial credence in any proposition A given any evidence E should plausibly equal $\mathrm{Conf}(A/E)$:

$$Cr_0(A/E) = \mathrm{Conf}(A/E).$$

I will call this principle *prior alignment*. Rudolf Carnap [1962] once suggested that prior alignment could serve to *define* the confirmation measure Conf. The idea is that judgements about the extent to which some evidence supports some hypothesis can be explicated as judgements about the initial credence one should assign to the hypothesis given the evidence. I have sympathies for this move, but for present purposes it does not matter whether prior alignment holds by definition or not.

In the presence of prior alignment, one might think that the conservative model of classical Bayesianism and the evidentialist model only come apart in situations where an agent loses information. The reasoning is simple. Above we saw that the credence function of a rational agent who always changes her beliefs by conditionalization will equal her initial credence function Cr_0 conditional on all her past and present evidence:

$$Cr_t(A) = Cr_0(A/E_1 \wedge \ldots \wedge E_t).$$

So, by prior alignment,

$$Cr_t(A) = \mathrm{Conf}(A/E_1 \wedge \ldots \wedge E_t).$$

If the agent's evidence is *cumulative* in the sense the evidence at later times entails all the earlier evidence – intuitively, if the agent never loses information – then E_t is equivalent

2 One could allow the confirmation measure Conf to vary with the agent or her external circumstances. One could also take Conf to measure a basic, agent-relative attitude of "confirmational commitments", approaching the models defended by Isaac Levi (e.g. [Levi 1980], [Levi 2010]) and Timothy Williamson ([Williamson 2000: chs. 9–10]). In the present paper, I will concentrate on the simplest form of evidentialism.

to $E_1 \wedge \ldots \wedge E_t$ and $\mathrm{Conf}(A/E_1 \wedge \ldots \wedge E_t)$ coincides with $\mathrm{Conf}(A/E_t)$. In that special case, then, the evidentialist principle and the conservative Bayesian model give the same verdict. Moreover, in other cases, where the agent has forgotten something she once learned, the evidentialist model seems preferable to the conservative model: the latter demands that the agent's credence should always reflect the impact of past evidence, which effectively rules out the possibility of information loss.

I will return to the problem of information loss in sections 7 and 8. Before that, I will discuss cases where conservative Bayesianism and evidentialism come apart even though the agent never loses information. This is possible because the reasoning just outlined rests on some simplifying assumptions. In particular, we have assumed that on the Bayesian model the agent only changes her beliefs by conditionalization. Once we take into account, for example, the different kind of change prompted by changes in the agent's own location, the argument no longer goes through. We have also assumed that there is a unique and determinate confirmation measure Conf, which makes evidentialism committed to the Uniqueness thesis ("given one's total evidence, there is a unique rational doxastic attitude one can take to any proposition" [White 2005: 455]). Let's begin by seeing what happens if we relax that assumption.

4 Fickle Frank

Many Bayesians are attracted to an intermediate position between radical subjectivism and radical objectivism: they agree that there are substantive constraints on initial credence beyond probabilistic coherence, but they do not agree that our norms of rationality allow for only one initial credence.

For example, consider the initial (or non-initial) credence we should assign to skeptical scenarios. Eric Schwitzgebel [2016] argues that it should be "around 0.1% to 1%, plus or minus an order of magnitude". That sounds plausible; it is hard to believe that our norms of rationality determine a unique and precise value – say, 0.23214%. If you are a little more cautious than me, giving slightly higher credence to skeptical scenarios, I should not fault you for being irrational.

For another example, consider the initial probability of physical theories. Most people agree that, all else equal, simple theories should have greater prior probability than complicated theories. But there are different aspects of simplicity, and different ways of measuring and balancing these aspects. Suppose one physical theory is more parsimonious by postulating fewer primitive quantities while another theory has mathematically simpler dynamical laws. Which of these has greater prior probability? If the two theories make the same predictions about the so-far observed aspects of the world, is there a uniquely rational attitude one may take towards them? Arguably not. It would be no sign of

irrationality if one theorist gives slightly higher credence to the parsimonious theory while another slightly favours the theory with the simpler dynamics.

Can evidentialists respect these permissivist judgements? They can. Let's replace the single confirmation measure Conf by a set CONF of measures, corresponding to the initial credence functions moderate Bayesians regard as acceptable. The permissivist principle of evidentialism now turns into the claim that if an agent's evidence at time t is E_t, then her credence function should coincide with *one of* the confirmation measures in CONF conditionalized on E_t; each of them is rationally permitted.

It is easy to see how permissive evidentialism and conservatism can come apart even if agents never lose information. The following scenario is loosely based on a case from [Hedden 2015b].[3]

> **Fickle Frank**. Frank is a physicist in a possible world where two theories A and B equally account for all physical phenomena that have so far been observed. Theory A is more parsimonious, theory B has simpler dynamical laws, and no theory does better than those two. Without receiving any relevant new evidence, Frank constantly changes his mind about which of the two theories he favours. In the morning, he bets on the truth of theory A, soon afterwards he tries to retract his bet because he has suddenly come to favour theory B, then he wants to make the original bet again because he has returned to his previous state of belief, and so on.

Assume Fred's evidence does not determinately favour one of the two theories over the other: relative to some measure in CONF, his evidence makes theory A more probable than theory B, relative to others, the evidence favours theory B. Permissive evidentialism then implies that there is nothing epistemically wrong with Fickle Frank. His beliefs are always proportioned to his evidence. On the other hand, Frank's fluctuating history of beliefs is clearly incompatible with the principle of doxastic conservatism. By assumption, whatever new evidence Frank receives throughout the day has no bearing on which of the two theories is true. Thus if in the morning Frank's credence in theory A was x, then his credence in theory A was also x conditional on whatever he was about to learn. By the principle of doxastic conservatism, Frank's later credence should therefore still have been x. That is also what would happen if Frank changed his beliefs by the rule of conditionalisation.

So conservatism, along with standard Bayesianism, regards Frank's fluctuating beliefs as a deviation from ideal rationality, permissive evidentialsm does not. Which side has it right? Our pre-theoretic judgement arguably sides with conservatism. When people

3 In Hedden's scenario, Frank switches between strong belief in different interpretations of quantum mechanics. I have changed the example because even permissivists have good reason to doubt that all these attitudes are equally in line with our (and Frank's) evidence.

change their mind, we assume that there should be a rational reason: new evidence, a re-assessment of previous evidence, a change in values, etc. A change in mind that has no such basis at all would strike us as problematic. Frank's fickle dispositions also lead to practical problems. If he bets in accordance with his beliefs, he will accept a series of bets whose net outcome is a guaranteed loss. More generally, it is hard to pursue long-term goals or plans if one's beliefs constantly change in unpredictable ways.

Committed evidentialists will dismiss these considerations as question-begging. According to evidentialism, what matters is only whether Frank's beliefs at any given point are adequately proportioned to his evidence. The fact that he incurs sure losses and is unable to follow plans is deplorable, on that view, but it does not reveal any epistemic shortcomings.

This dialectical situation will arise in all the cases I will discuss. No doubt evidentialism is an internally consistent theory. None of the cases I will discuss therefore casts doubt on evidentialism *from the perspective of evidentialism*. But the way to evaluate conflicting normative proposals is not to assume one of the proposals as true and reject all objections as question-begging. Rather, we should ask to what extent the different proposals are in line with our "pre-theoretic" (but considered) normative judgements, how well they cohere with other normative and non-normative commitments, and whether they fit our best ideas about the grounds of the relevant kind of normativity. I will offer a few general considerations along those lines in section 8. Until then, I mainly want to convince you that evidentialism and conservatism can pull in different directions, and that where they do, there are reasons to side with conservatism.

The present case rests on the permissivist assumption that there is often a range of eligible credence functions. Many evidentialists reject this assumption. In line with the Uniqueness thesis, they hold that our evidence always determines a unique rational state of belief. Some try to make this more palatable by allowing for "imprecise" states, so that the attitude Frank ought to take towards the two theories is not a precise and determinate probability such as 0.43279, but a (precise and determinate) probability interval such as [0.39012, 0.48852], perhaps corresponding to the different verdicts of the confirmation measures in Conf.[4]

Uniqueness avoids the clash between evidentialism and conservatism in cases like *Fickle Frank*. But it puts a lot of weight on the concept of objective support: if tenable forms of evidentialism are committed to a unique and determinate measure of evidential support,

4 In the present paper, I will generally assume that confirmation measures and rational credence functions take precise values, and that rational agents have a single credence function. This is only for the sake of simplicity – all the issues I discuss would also arise if we allowed for imprecise credence. There is nonetheless a connection between my present topic and the debate over imprecise credence that might be worth exploring: imprecision is usually motivated by considerations of evidential support (e.g. [Joyce 2005]), while precision is often supported by diachronic considerations, especially concerning sequential choice (e.g. [Elga 2010]).

then setting aside any qualms about that measure is a tall order.

Moreover, while the Uniqueness thesis allows evidentialists to identify *a* problem with Fickle Frank, it arguably doesn't identify the *right* problem. When Frank changes his mind without receiving relevant evidence, Uniqueness implies that either his earlier or his later state is not in line with his evidence. But intuitively, that's not what is wrong with Frank. What is wrong is that his beliefs constantly change for no reason whatsoever, which gives rise to his erratic and costly behaviour. The problem lies in the *difference* between his earlier and later belief state, not in either of those states taken in isolation. If Frank had two colleagues, one of whom consistently favoured theory A and the other theory B, there would be something wrong with Frank that is not wrong with his colleagues.

I do not have much more to say in defense of these claims, so let me move on to the next scenario. For the rest of the paper, I will stack the deck in favour of evidentialism by assuming that there is a unique and determinate objective confirmation measure. As we will see, this still leaves room for a variety of cases in which evidentialism and conservatism pull in opposite directions.

5 Dr. Evil

When we wake up at night and wonder what time it is, the object of our uncertainty appears to be an "essentially indexical" or "self-locating" proposition, a kind of proposition that takes different truth-values at different times and places (see e.g. [Lewis 1979]). Including such propositions among the objects of credence makes no big difference to evidentialism, but it raises a minor challenge for Bayesian conservatism, for it then becomes implausible that rational credence simply evolves by conditionalization.

To illustrate, suppose before going to sleep an agent is fairly confident that it is 11pm and that she will wake up at 6am. Upon awakening, she then ought to be confident that it is 6am. By conditionalization, her new credence would have to equal her old credence conditional on her new evidence. What is that new evidence? Perhaps she has a diffuse sensation as of waking up from several hours of sleep. However, her 11pm credence conditionalized on the hypothesis that she has that sensation hardly equals her rational 6am credence. At 11pm, we can assume the agent was quite confident that she does not currently have a sensation of awakening. It is not clear what she should have believed, at 11pm, conditional on the incredible hypothesis that she does (right now) have such sensations. She might well have concluded that she has a strange neurological disorder that dissociates her experiences from her awareness. That is obviously not what she judged upon awakening in the morning.

The problem is widely recognized, and several answers have been put forward. Some have suggested moving to a form of evidentialism, at least with respect to self-locating

propositions (see e.g. [Halpern 2006], [Briggs 2010]). On that view, an agent's new credence in self-locating propositions is determined by her new evidence alone, irrespective of her previous credence in self-locating propositions. More conservative proposals have generally added a further operation to conditionalization that governs the evolution of self-locating beliefs. To borrow terminology of [Katsuno and Mendelzon 1991], conditionalization is a rule for *revising* a belief state in the light of new information; another rule is required for *updating* a belief state to keep track of changes in the agent's location in space and time (see e.g. [Kim 2009], [Meacham 2010], [Schulz 2010], [Bradley 2011b], [Schwarz 2012]).[5]

The precise details of the new rule will not be too important in what follows, but let me sketch the basic idea. Return to our awakening agent. We have assumed that at 11pm, the agent was confident that she was about to sleep until 6am. On the present approach, this is enough to settle that when she finds herself awakening she ought to believe that it is 6am, even without looking at her clock or receiving other relevant information. Given her background belief that she would wake up seven hours later, her previous belief that it is 11pm should *update* to a belief that it is 6am. If before falling asleep the agent had been unsure whether she would wake up at 6am or 7am, the update step would leave the agent uncertain whether it is now 6am or 7am.

In [Schwarz 2012] and [Schwarz 2015], it is shown that the update+revision process inherits many characteristic features of conditionalization once self-locating propositions are taken into account. For example, agents are vulnerable to diachronic Dutch Books if and only if their beliefs do not evolve in accordance with update+revision.

Along with the norms of belief dynamics, we must also adjust the principle of conservatism. Suppose upon awakening an agent's evidence E_t includes the (self-locating) information that dawn is approaching. According to the principle of conservatism as formulated above, the agent should retain her previous credence in any proposition A – say, that her clock is showing the correct time – provided that her previous credence in A was equal to her previous credence in A given E_t. But clearly we shouldn't consider the agent's 11pm credence in her clock being right conditional on the hypothesis that dawn is approaching. We have to take into account the fact that E_t changes its truth-value between 11pm and 6am. Arguably what matters is the agent's 11pm credence in her clock being right conditional on the hypothesis that dawn *will* approach when she wakes up. Even better, since the clock also might stop (or start) working during her sleep, we should consider the 11pm credence in the hypothesis that her clock *will* be right given that dawn *will* approach. In general, I suggest that the adjusted principle of conservatism should say that if before t the agent treated the hypothesis that A would be true as

5 A minority of Bayesians – notably Robert Stalnaker (e.g. [Stalnaker 2008: ch.3], [Stalnaker 2014: ch.5]) – have resisted the idea that objects of credence can change their truth-value over time. However, even on those accounts credences cannot evolve simply by conditionalization. On Stalnaker's account, the cases I am about to discuss require non-trivial steps of "recalibrating" the posterior credences.

independent of the hypothesis that E_t would be true, then upon subsequently finding that E_t is true, her credence in A should equal her previous credence that A would be true.[6] If that sounds complicated, the following much simpler special case will mostly be sufficient for purposes of the present paper: if A is certain not to change its truth-value, and before learning E_t the agent was already certain that she would learn E_t, then learning E_t should not affect her credence in A.

It is easy to show that the update+revision norm implies the new principle of conservatism, just as conditionalization implied the old principle. Once again, we therefore get the corollary that agents who violate the new principle of conservatism are vulnerable to diachronic Dutch Books.

With all that set up, let's reconsider a well-known scenario from [Elga 2004].

> **Dr. Evil.** In a battlestation on the moon, Dr. Evil receives a letter informing him that a perfect duplicate of Dr. Evil, "Dup", has just been created on Earth. Dup inhabites a duplicate of Dr. Evil's battlestation, reads a duplicate of the letter addressed to Dr. Evil and overall has experiences indistinguishable from those of Dr. Evil. For some reason, Dr. Evil is rationally certain that the letter is true.

Elga argues that Dr. Evil should become 50% confident that he is Dup. Elga's argument assumes evidentialism; as we will see, conservatism instead suggests that Dr. Evil should remain confident that he is Dr. Evil. Let me begin with the evidentialist side.

Elga's own argument is rather complicated; for the present discussion a simplified version will suffice. We need two premises to show that Elga's judgement is implied by evidentialism. The first is that Dup and Dr. Evil have the same evidence. This could be denied. For example, one could argue that whatever an agent remembers is part of their evidence and that only Dr. Evil remembers having flown to the moon; Dup has "quasi-memories" of such a flight, but they do not count as evidence. Now before we decide whether something does or doesn't count as evidence, it is advisable to clarify what role an agent's evidence is meant to play. In the present context, the central role of evidence is given by the principle of evidentialism: evidence is something that determines rational belief in the manner expressed by that principle. Evidentialists may now debate whether that something is the same between Dup and Dr. Evil. I will assume we are dealing with a version of evidentialism on which Dup and Dr. Evil count as having the same evidence.

The second premise is that possibilities that differ merely with respect to matters of self-location should have equal prior probability. This, too, could be questioned, but we will see that it is not really needed to establish the contrast between evidentialism and conservatism.

6 In the notation of [Schwarz 2012]: if $Cr_{t-1}(\succ A/\succ E_t) = Cr_{t-1}(\succ A)$ then $Cr_t(A) = Cr_{t-1}(\succ A)$.

Given those two premises, the evidentialist argument is straightforward. Let E be Dr. Evil's evidence when he has read the letter. Since Dr. Evil is rationally certain that the letter is true, E rules out all possible worlds in which Dr. Evil does not have a duplicate on Earth. By premise 2, conditional on any of these worlds, the self-locating possibility of being Dr. Evil has equal prior probability as the self-locating possibility of being Dup. Moreover, by premise 1, Dup and Dr. Evil have the same evidence in all E-worlds. It follows that among E-worlds, Dr. Evil's evidence never excludes a Dup possibility without also excluding the corresponding Dr. Evil possibility and vice versa. Hence conditional on E, being Dr. Evil and being Dup have equal probability.

What if Dr. Evil obeys conservative norms? The answer then depends on his earlier beliefs, which Elga does not tell us. Let's flesh out the scenario as follows. Long before Dr. Evil received the letter, a reliable spy informed him about the plans on Earth to build a perfect duplicate of Dr. Evil and his surroundings. At that point, before the duplicate was built, Dr. Evil had no special reason to doubt that he is Dr. Evil, located on the moon. Let's say he rationally gave credence 0.99 to that assumption. Then time passed, the duplicate was built, and the letter arrived. What happened to Dr. Evil's belief state if it went through the process of update and revision? We can assume that for the whole time, Dr. Evil was certain that *if* he is presently on the moon, then he will continue to be on the moon in the near future. (He has no plans to travel, and is certain that he won't be hijacked.) Consequently, the update step never moves any probability away from the assumption that he is on the moon. It does however move more and more probability to the hypothesis that he has a duplicate on Earth, as he expects the relevant plans on Earth to materialize. As long as Dr. Evil receives no other relevant information, conditionalizing his updated credences on new information also won't move any probability away from the assumption that he is on the moon. Eventually the arrival of the letter informs him that the constructions on Earth have finished. At that point, Dr. Evil may well have been fairly confident already that a duplicate had been created on Earth, so the letter carried little news. Indeed, elementary Bayesian reasoning shows that whatever new evidence Dr. Evil receives through the letter could not raise his credence in being Dup, for that would require that his evidence is *more likely* conditional on being Dup than conditional on being Dr. Evil, which we can certainly rule out. So Dr. Evil will remain 99% confident that he is Dr. Evil.

Instead of applying update and revision, we can also apply the (adjusted) principle of doxastic conservatism. Assume that Dr. Evil receives no relevant information between the news from the spy and the arrival of the letter. At the beginning of that period, Dr. Evil was rationally confident that he is on the moon. How confident was he that he will still be on the moon later conditional on the assumption that he would then receive the letter? Strongly confident. By the adjusted principle of doxastic conservatism, receiving the letter therefore should not affect Dr. Evil's credence that he is on the moon.

The case is even more obvious if we assume that the spy already revealed to Dr. Evil that he was going to receive the letter and how long it would take until the letter would arrive. The arrival of the letter then presented no news at all to Dr. Evil. So we can apply the simplified principle of conservatism, according to which, if you learn something of which you were certain beforehand that you would learn it, you should not change your credence in propositions that are certain not to change their truth-value. By that principle, Dr. Evil should remain confident that he is on the moon.

Elga never mentioned a spy. You might therefore worry whether the present considerations carry over to Elga's original, spy-free scenario. It doesn't matter. For the evidentialist argument above clearly carries over to the present scenario: if the two premises are true in the spy-free scenario then they are also true in the spy scenario. So we have what we wanted – a case where evidentialism and conservatism come apart.

Strictly speaking, as I mentioned above, what comes apart from conservatism here is any brand of evidentialism on which Dr. Evil and Dup count as having the same evidence. It is tempting to think that analogous cases could be construed to target other forms of evidentialism, by extending the construction of Dup until sameness of evidence is reached. I will not explore the matter any further here, since most actual evidentialists arguably accept the "sameness of evidence" premise in Elga's scenario.

I can now explain why we don't need Elga's second premise, his principle of "self-locating indifference". Suppose we maintain that on the contrary, skeptical scenarios always have lower evidential probability than corresponding non-skeptical scenarios, even when the two are located in the same possible world. Dup possibilities should then have lower probability than Dr. Evil possibilities. But presumably they do not have probability zero. As skeptical scenarios go, the Dup scenario is not even especially radical (and it could be made even less radical if desired). On the other hand, if the details are spelled out appropriately, conservatism allows Dr. Evil to give arbitrarily low credence to being Dup. So evidentialism and conservatism still come apart.

Who has it right? Elga expresses a strong intuition that Dr. Evil should become unsure whether he is on the moon: his evidence doesn't tell! I would press the contrary intuition: Dr. Evil should not drastically revise his beliefs about his location in response to completely unsurprising evidence that has no bearing on where he is!

Unlike in the case of Fickle Fred, I am not sure here which side is closer to untutored judgement. General criteria of diachronic rationality unsurprisingly favour the conservative verdict. Thus obeying evidentialism makes Dr. Evil vulnerable to a straightforward diachronic Dutch Book. Observe also that obeying evidentialism comes with a predictable high cost in accuracy. If Dr. Evil obeys conservative norms, he will end up giving high credence to the true propositions that he is Dr. Evil on the moon, that he once flew there, and so on – and he will do so not as a matter of luck. By contrast, if he obeys evidentialism, his credence in true propositions will be significantly reduced. To the

extent that reliable accuracy is an epistemic goal, this seems to favour the conservative approach.[7]

6 Duplication machines

The next scenario I want to look at involves an agent who gives some credence to the hypothesis that he will undergo a process of personal fission or duplication. This kind of scenario may appear far-fetched, but it is worth remembering that the main realist understanding of quantum physics, the Everett interpretation, arguably implies that personal fission happens all the time (see e.g. [Wallace 2012]). Those of us who give non-zero credence to the Everett interpretation therefore constantly find ourselves in scenarios in which personal fission is a live epistemic possibility.

Before I turn to my main example, I want to spend a minute to think about how conservative norms apply in a case of fission. Imagine an agent, call him Fred, who knowingly enters a duplication machine that will (say) cut him apart lengthwise and then fuse each half with a perfect copy of the other half, producing two perfect copies of the original person. Let's not worry about the metaphysical question which of the two persons emerging from the machine (which I will call Fred's *successors*), if any, is identical to the original Fred. Rather, let's ask how Fred's belief state should be updated conservatively so as to produce reasonable belief states in the two successors.[8]

For example, if before entering the machine, Fred was confident that penguins eat fish, a conservative update mechanism would arguably preserve that belief: the two successors should still be confident that penguins eat fish. But what about self-locating beliefs? Suppose Fred knows that one of his successors will emerge on the left, the other on the right. Before the successors receive any new information about their whereabouts, what should they believe about where they are? It seems reasonable to say that they should give equal credence to being on the left and being on the right. After all, neither the successors nor Fred ever acquired any information that would favour one side over the other.

7 The accuracy of Dr. Evil's belief state is compensated by Dup's: if Dr. Evil is confident that he is on the moon, then Dup, who is programmed to duplicate Dr. Evil, will also be confident that he is on the moon and thus have highly inaccurate beliefs. But the idea that belief aims at truth surely isn't reasonably cashed out in a quasi-utilitarian fashion where the goal is to maximize the overall accuracy in the world.

8 Brian Hedden [2015b: 456f.] claims that in order to apply diachronic norms we would first have to settle matters of personal identity. I'm more inclined to see the dependence go the other way: one of the criteria that guide us in treating an agent at a later time as the same person as an agent at an earlier time is that the transition from the earlier belief state to the later belief state does not deviate too much from the norms of diachronic rationality. In any case, I don't see why we couldn't ask how beliefs should evolve through a process of fission while remaining neutral on matters of personal identity.

In thinking about such cases, it can be useful to adopt David Lewis's [1976] model of fission on which Fred is actually two co-located persons even before entering the duplication machine. One of these, call him *Fred*$_L$, will emerge on the left, the other, *Fred*$_R$, on the right. Since Fred$_L$ and Fred$_R$ start out as perfect (intrinsic and extrinsic) duplicates, they presumably must have the same degrees of belief. Furthermore, all the evidence they receive before the fission event is neutral on whether they are Fred$_L$ or Fred$_R$. Finally, neither of these hypotheses has greater prior plausibility than the other. It is therefore plausible that the two Freds should give equal credence to being Fred$_L$ and being Fred$_R$ both before and after the fission, until they start receiving evidence that allows them to tell the two possibilities apart.

Now let's turn to the following scenario, which does not involve an actual episode of fission.

> **The broken duplication machine.** Fred has bought a duplication machine at a discount from a series in which 50 percent of all machines are broken. If Fred's machine works, it will turn Fred into two identical copies of himself, one emerging on the left, the other on the right. If Fred's machine is broken, he will emerge unchanged and unduplicated either on the left or on the right, but he can't predict where. Fred enters his machine and finds himself emerge on the left. In fact, Fred's machine is broken and no duplication event has occurred, but his experiences do not reveal this to him.

What credence should Fred give to the hypothesis that his machine works? I will argue that by conservative lights Fred's credence should be 1/2 while evidentialism says it should be 2/3.

The basic conservative argument is simple. Before he entered the machine, Fred should clearly have been 50% confident that his machine works, given his knowledge that the machine comes from a series in which 50% of all machines are broken. Moreover, whatever Fred learns upon exiting the machine sheds no new light on whether his machine works or not. Hence Fred should remain 50% confident that his machine works.

The argument can be spelled out more carefully by the update+revision rule. Again, Lewis's model of fission helps to streamline the application. On Lewis's model, Fred's credence is initially divided between four self-locating possibilities, which Fred might express as follows:

1. The machine works and I will emerge on the left;
2. The machine works and I will emerge on the right;
3. The machine does not work and I will emerge on the left;
4. The machine does not work and I will emerge on the right.

When he enters the machine, Fred should give credence 1/4 to each of these hypotheses. The update step will then produce a state in which Fred gives credence 1/4 to each of the following:

1. The machine works and I have emerged on the left;
2. The machine works and I have emerged on the right;
3. The machine does not work and I have emerged on the left;
4. The machine does not work and I have emerged on the right.

Observing that he has emerged on the left, Fred can exclude possibilities 2 and 4, leaving him with 50% credence in possibility 1 and 50% in possibility 3. So Fred will still be 50% confident that his machine works.[9]

We may also directly apply the principle of conservatism. Let's assume that upon leaving the machine, Fred keeps his eyes closed for a moment, so that he does not receive any relevant new information at all. Let's also assume that Fred knows this in advance. We can then use the simplified principle according to which Fred's credence in the hypothesis that his machine works should not change if he knew in advance what he would learn upon exiting. Before opening his eyes, Fred should therefore be 50% confident that his machine works. Moreover, both conditional on the hypothesis that his machine works and conditional on the hypothesis that the machine doesn't work, Fred should now be 50% confident that he is on the left. So when he finally opens his eyes and sees that he is on the left, he learns nothing new about whether his machine works.

Yet another way to see that diachronic norms should support the answer 1/2 is to note that emerging on the left and emerging on the right are completely symmetrical. If Fred's credence in the functioning of his machine should increase (or decrease) when he emerges on the left, then it would also have had to increase (decrease) had he emerged on the right. His credence should have gone up (down) no matter what he learns. That seems problematic; among other things, it seems to make him vulnerable to diachronic Dutch Books.[10]

So conservatism says that Fred should remain neutral on whether his machine works. Turning to evidentialism, let's go through Fred's evidence when he emerges from the machine and finds himself on the left. We can assume that Fred remembers that 50 percent of the relevant machines are broken, which makes it 50 percent probable that Fred's machine is broken. In addition, Fred knows that he has just emerged on the left. That is, he knows that either his machine is broken and the single original Fred has emerged on the left, or the machine works and one of Fred's duplicate successors

9 See [Schwarz 2015] for how to apply update+revision without assuming Lewis's model, and [Greaves 2007] for a superficially different dynamical model that gives the same result.

10 The evaluation of diachronic Dutch Books in cases of fission is not entirely trivial as it is not obvious how pre-fission and post-fission payoffs should be added up. For reasons of space I will not enter this discussion.

has emerged on the left. Let's abbreviate this information as 'some Fred has emerged on the left', using 'Fred' as a count noun that applies both to the original Fred and to his duplicate successors. Supposing the machine works, it is certain that some Fred would emerge on the left. Supposing the machine is broken, there is only a 50 percent probability of that outcome. The information that some Fred has emerged on the left therefore supports the hypothesis that his machine works.

Let me go through this line of thought more slowly. (Feel free to skim the next two pages if you're already persuaded.) Let *Works* be the proposition that Fred's machine works, let B be the background information that 50 percent of the relevant machines work and that Fred has recently entered his machine, and let *Fred Left* be the proposition that some Fred has emerged on the left. Then

$$\begin{aligned} \text{Conf}(\textit{Works}/B) &= 1/2, \\ \text{Conf}(\textit{Fred Left}/\textit{Works} \,\&\, B) &= 1, \\ \text{Conf}(\textit{Fred Left}/\neg \textit{Works} \,\&\, B) &= 1/2. \end{aligned}$$

By Bayes' Theorem[11], it follows that

$$\text{Conf}(\textit{Works}/\textit{Fred Left} \,\&\, B) = 2/3.$$

So if Fred's (relevant) evidence is *Fred Left* & B, then the evidence supports the hypothesis that his machine works to degree 2/3.

For another argument that Fred's evidence supports *Works* over $\neg$*Works*, imagine that Fred's friend Ted is watching Fred test his machine. Ted is positioned so that he can only see who emerges on the left-hand side. He knows everything that Fred knows. He knows that 50 percent of the relevant machines are broken, and that Fred just entered this particular machine. At that point, Ted's credence in the hypothesis that Fred's machine works is 1/2. Ted also knows that if the machine works, then some Fred is bound to appear on the left, whereas if the machine is broken then it is just as likely as not that some Fred will appear. So when Ted sees Fred appear, his credence in the hypothesis that Fred's machine works should increase.

Unlike the case of Fred, the case of Ted should be uncontroversial: even conservatives will agree that Ted's new credence in *Works* should be 2/3. (Note that Ted's credence would not have increased no matter what.) But Fred and Ted do seem to have all the same (relevant) evidence. We can even imagine that they are free to communicate and share all their evidence. Neither of them would thereby gain any relevant news. So if evidence can in principle be shared between agents, then Fred's evidence must also support the hypothesis that his machine works to degree 2/3.

11 Bayes' Theorem states that for any probability measure P and propositions H and E, $P(H/E) = P(E/H)P(H)/P(E)$.

It might be objected that there is a special kind of self-locating evidence that cannot possibly be shared and that we neglected in the above application of Bayes' Theorem: Fred knows not only that *some Fred* has emerged on the left, but also that *he himself* has emerged on the left. Let *I Left* express this further information. Now while it is true that

$$\begin{aligned}\mathrm{Conf}(\textit{Fred Left}/\textit{Works} \,\&\, B) &= 1,\\ \mathrm{Conf}(\textit{Fred Left}/\neg \textit{Works} \,\&\, B) &= 1/2,\end{aligned}$$

the objector might argue that

$$\begin{aligned}\mathrm{Conf}(\textit{I Left}/\textit{Works} \,\&\, B) &= 1/2,\\ \mathrm{Conf}(\textit{I Left}/\neg \textit{Works} \,\&\, B) &= 1/2,\end{aligned}$$

In contrast to *Fred Left*, *I Left* therefore would not raise the probability of *Works*.

In response, let's grant that there are such self-locating propositions that can't possibly be shared. But it is not plausible that

$$\mathrm{Conf}(\textit{I Left}/\textit{Works} \,\&\, B) = 1/2.$$

Recall that we are talking about objective, evidential probabilities. To what extent does the hypothesis that Fred's machine works together with the background information about the machine's origin and Fred's recent entering of the machine support the self-locating hypothesis 'I have emerged on the left'? Surely not to degree 1/2. At most, we have

$$\mathrm{Conf}(\textit{I Left}/\textit{Works} \,\&\, B \,\&\, \textit{I Fred}) = 1/2,$$

where *I Fred* is the self-locating proposition that might be expressed by 'I am a Fred'. The evidential probability of *I Fred* conditional on *Works* & B is not 1: otherwise Ted, who also knows B, should be certain that if Fred's machine works then *he* (Ted) has emerged either on the left or on the right. So if we take into account the self-locating information *I Left*, we also have to take into account *I Fred*, which we may assume is part of Fred's total evidence. But *I Fred* increases the evidential probability of *Works*. The reason is that there are more Freds in worlds where the machine works than in worlds where it is broken. For example, let H_6 be the hypothesis that there are in total six individuals in the world at the relevant time, including Ted and however many Freds have emerged from the machine. Since $B \,\&\, H_6$ is neutral on issues of self-locating, the six self-locating possibilities are presumably equally supported by $B \,\&\, H_6$, as well as by $B \,\&\, H_6$ & *Works*. So

$$\begin{aligned}\mathrm{Conf}(\textit{I Fred}/B \,\&\, H_6 \,\&\, \textit{Works}) &= 1/3,\\ \mathrm{Conf}(\textit{I Fred}/B \,\&\, H_6 \,\&\, \neg \textit{Works}) &= 1/6,\end{aligned}$$

and so, by Bayes' Theorem,

$$\text{Conf}(\textit{Works}/B \,\&\, H_6 \,\&\, I\ \textit{Fred}) = \frac{1/3 \cdot 1/2}{1/3 \cdot 1/2 + 1/6 \cdot 1/2} = 2/3.$$

The argument obviously generalizes to other hypotheses H_n about the number of individuals in the world, and to cases where some of the extra individuals deserve lower prior probability.[12] Thus by the law of total probability,

$$\text{Conf}(\textit{Works}/B \,\&\, I\ \textit{Fred}) = 2/3.$$

Moreover, the background information B & I *Fred* lends equal support to I *Left* and I *Right*, both on the assumption that the machine works and on the assumption that it is broken:

$$\text{Conf}(I\ \textit{Left}/\textit{Works} \,\&\, B \,\&\, I\ \textit{Fred}) = \text{Conf}(I\ \textit{Left}/\neg\textit{Works} \,\&\, B \,\&\, I\ \textit{Fred}) = 1/2.$$

It follows (again by Bayes' Theorem) that

$$\text{Conf}(\textit{Works}/B \,\&\, I\ \textit{Fred} \,\&\, I\ \textit{Left}) = 2/3.$$

The upshot is that taking into account Fred's essentially self-locating information makes no difference: the evidential probability of *Works* given Fred's total evidence B & I *Fred* & I *Left* is still 2/3.

Let me give one more argument for that conclusion.[13] Consider a variant scenario in which the only problem with a broken duplication machine is that one of the two successors it produces will have a damaged retina and thus be blind. (Fred knows this, but he can't predict whether the blind successor would emerge on the left or on the right.) Assume also that anyone who emerges from the machine keeps their eyes closed for a moment. Since Fred's machine is broken, it produces a blind successor – on the right, say – and a non-blind successor on the left. Before opening their eyes, both successors should plausibly assign equal credence to *Works* and ¬*Works*; they certainly don't seem to have any evidence that supports one hypothesis over the other. Moreover, among ¬*Works* possibilities, their credence should be equally divided between the hypothesis that they are blind and the hypothesis that they can see. When the successor on the left then discovers that he can see, he can rule out all ¬*Works* & *Blind* possibilities but none of the previously open *Works* possibilities; so his credence in *Works* should be 2/3. Learning I *Left* has no further effect because his credence among the remaining ¬*Works* & ¬*Blind*

12 It might not generalize to the hypothesis that there are infinitely many individuals, but we can ignore that complication by adding to B the information that the universe is finite.

13 The argument resembles a well-known argument in the Sleeping Beauty debate, to which I will turn in the next section.

possibilities should be evenly divided between *I Left* and *I Right*, and similarly for his credence among the remaining *Works* possibilities.

Now compare the evidence available to the left-hand successor in the variant scenario with Fred's evidence in the original scenario. The evidence is not exactly the same, but with respect to the question whether Fred's machine works it is completely on a par. *From an evidentialist perspective*, it follows that the credence in *Works* should be the same in either case.[14]

To sum up, when Fred emerges from his machine, conservatism says that his credence in the hypothesis that the machine works should remain at 1/2, while evidentialism says it should increase to 2/3. My own intuitions again side with conservatism, and there is evidence that I'm not alone. Scenarios much like the present one have received considerable attention in the literature on Everettian Quantum Mechanics (see e.g. [Greaves 2007], [Lewis 2007], [Papineau and Durà-Vilà 2009], [Titelbaum 2013: ch.11]), and it is widely agreed that the answer I classified as conservative is correct. Most of the debate instead concentrates on the interpretation of future-directed probabilities in fission scenarios and on whether there are parallels between the present type of case and the Sleeping Beauty problem.

Which brings me to the next and last scenario I want to discuss.

7 Sleeping Beauty

> **Sleeping Beauty.** On Sunday night, after Sleeping Beauty has gone to sleep, a fair coin is tossed. If the coin lands tails, Beauty's state of mind before her awakening on Tuesday morning will be reset to her state of mind before her awakening on Monday morning. If the coin lands heads, she will instead be made to sleep all through Tuesday. Beauty knows of this arrangement.

When she wakes up on Monday, what credence should Beauty give to the hypothesis that the coin has landed heads? "Halfers" say it should be 1/2, "thirders" say 1/3. After almost 20 years of debate there is still no sign of agreement between the two sides. It is getting increasingly unlikely that one party has simply made a calculation mistake. A more charitable hypothesis is that, like Newcomb's problem, the Sleeping Beauty problem brings to light a deeper normative disagreement. I want to suggest that the underlying

14 To emphasize, the argument here is not that the two scenarios are intuitively analogous and therefore should have the same answer. In fact, from a conservative perspective, the left-hand successor's credence in *Works* should be 2/3 in the variant scenario whereas Fred's credence in the original scenario should be 1/2, so the two scenarios are definitely not analogous. The present argument is that any difference between the evidence in the two scenarios has no bearing on whether the machine works, from which it follows not "by analogy" but by the principle of evidentialism that the credences should be the same.

disagreement is the disagreement between evidentialism and conservatism: conservatism supports halfing, evidentialism thirding.

A powerful evidentialist argument for thirding, outlined in [Piccione and Rubinstein 1997], [Dorr 2002], [Arntzenius 2003], [Horgan 2004], [Horgan 2008], and [Horgan and Mahtani 2013], goes as follows. When Beauty wakes up on Monday, her evidence with respect to the coin toss consists in (1) her knowledge of the experimental setup, (2) her experience of being awake, and (3) the fact that she has no memories from later than Sunday. By itself, (1) lends equal support to the four combinations *Heads & Monday*, *Heads & Tuesday*, *Tails & Monday*, *Tails & Tuesday*; (3) rules out any further possibilities (such as *Heads & Wednesday*), and (2) excludes *Heads & Tuesday*. The remaining possibilities should therefore have probability 1/3 each. Assuming that Beauty should proportion her beliefs to the evidence, it follows that her credence in *Heads* should be 1/3.

A possible weakness in this argument is the assumption that possibilities in which one isn't awake (such as *Heads & Tuesday*) should have significant prior probability (see [Pust 2008]). We can bypass this worry by modifying the scenario so that Beauty wakes up on Tuesday even if the coin lands heads, in this case without having her belief state reset. It is then highly plausible that Beauty's knowledge of the setup confers equal evidential probability on all four combinations of *Heads* and *Tails* with *Monday* and *Tuesday*, one of which (*Heads & Tuesday*) is then ruled out by the fact that Beauty has no memories from later than Sunday. Conservatism supports halfing both in the original and in the modified scenario, so for present purposes the modification is harmless.

The conservative case for halfing is a bit more complicated. To be sure, there is the straightforward conservative argument, discussed for example in [Elga 2000], [Lewis 2001], and [Bradley 2011a]. The argument is that (1) Beauty's credence in *Heads* should have been 1/2 on Sunday, and (2) whatever Beauty learns when she wakes up on Monday has no bearing on whether the coin landed heads or tails. The second premise – which some thirders have challenged – could be substantiated by the principle of conservatism. Consider all the information Beauty has on Monday morning: that she is awake, that she has no memories from later than Sunday, that it is raining (say), and so on. Before Beauty went to sleep on Sunday, what should have been her credence in *Heads* conditional on the assumption that upon awakening she would have all that information? The answer is plausibly 1/2. So whatever Beauty learns on Monday is neutral on the outcome of the coin toss, as judged by her previous beliefs. By the principle of conservatism, her new credence in heads should therefore still be 1/2.

Now for the complication. Before she went to sleep on Sunday, Beauty knew that her next awakening would take place on Monday. The principle of conservatism (as well as the update+revision rule) therefore seems to imply that she ought to be certain upon awakening that it is Monday. In fact, she ought to be certain that it is Monday even if

she knew in advance that the coin would land tails. That seems wrong. It seems wrong not just because Beauty's evidence is neutral between *Tails & Monday* and *Tails & Tuesday* – which conservatives might well dismiss as irrelevant. It also seems wrong because Beauty will then be certain that it is Monday both on Monday and on Tuesday, and that does not look like an optimal way to deal with Beauty's predicament.

The deeper problem here is that the constraints of the story make it impossible for Beauty to follow the norms of diachronic rationality, at least if the coin lands tails: she will inevitably violate those norms in the transition from Monday to Tuesday. For example, whatever Beauty learns on Monday she will no longer know on Tuesday, even it does not contradict any new evidence. As an ideal conservative agent, Beauty would retain what she learns on Monday. More importantly, she would be certain on Monday that it is Monday and on Tuesday that it is Tuesday. The constraints of the scenario make that impossible. Arguably this fact is relevant already to the transition from Sunday to Monday.

Again, let's suppose for a moment that the coin is certain to land tails. How should Beauty's belief state evolve from Sunday night to Monday morning, given that the resulting state will also be her state on Tuesday morning? If she becomes certain that it is Monday, as the above naive application of conservative norms suggests, she will have very accurate beliefs on Monday and very inaccurate beliefs on Tuesday. But why should an optimal update privilege the Monday state?

Notice the parallels to cases of fission. Due to the mental reset on Monday night, the "epistemic predecessor" of Beauty's Tuesday morning state is not her Monday evening state but her state on Sunday evening. Her Sunday state therefore has two epistemic successors: one on Monday, one on Tuesday. I do not mean that Beauty actually undergoes a process of personal fission. Whether she does arguably depends on details of the scenario that are usually left open. Rather, my point is that if we ask what update mechanism we would ideally implement for agents in Beauty's predicament, the question is completely analogous to the corresponding question in cases of fission: we are looking for a mechanism that transforms the earlier belief state into a new belief state that will be instantiated at two different locations, and we would like to make both of these as accurate as possible. Just as it would not be ideal in a case of fission if both fission products became certain that they are emerging on the left, so it would not be ideal for Beauty if she became certain both on Monday and on Tuesday that it is Monday. Instead, the optimal update should plausibly leave her indifferent between *Monday* and *Tuesday*.

Now return to the original story where Beauty does not know that the coin lands tails. This makes the case analogous to one of merely possible fission. Accordingly, an optimal conservative update should not make Beauty indifferent between *Monday* and *Tuesday*. After all, if the coin lands heads, then the only epistemic successor to her Sunday state

will be located on Monday. The greater the probability of heads, the more the update should therefore favour the hypothesis that it is Monday. At the limit, if it is certain that the coin lands heads, Beauty should wake up certain that it is Monday. With the 50% chance of heads, as in the original story, Beauty's credence in *Monday* should plausibly be right between 1/2 and 1, at 3/4. This can be confirmed by applying the update+revision rule for cases of possible fission. (I spare you the details.) Beauty's credence in *Heads*, of course, will remain unchanged at 1/2, just like Fred's credence in the hypothesis that his machine is broken.

The key point in this argument for halfing (which has not yet been made in the literature) is that one cannot simply apply standard norms of diachronic rationality if there is an exogenous threat of diachronic irrationality. We need a fallback norm that takes the threat into account. For Sleeping Beauty, the relevant norm will assimilate Beauty's predicament to a case of possible fission. For the crucial feature of (binary) fission is that the updated credence will be instantiated twice over, and this is precisely what happens to Beauty if the coin lands tails.

A large number of arguments have been given for either halfing or thirding besides the two arguments I have discussed. Understanding the two positions as motivated by different views on diachronic rationality – conservatism in the case of halfing and evidentialism in the case of thirding – helps to make sense of these arguments. I will not go through every single argument in the literature, but I will illustrate the point with two more examples.[15]

In defence of thirding, [Piccione and Rubinstein 1997], [Dorr 2002] and [Arntzenius 2003] (independently?) present the following variation of *Sleeping Beauty*, which resembles the variation of *The broken duplication machine* discussed in the previous section. This time, Beauty is awakened on Tuesday no matter the outcome of the coin toss, with her mental state reset to her Monday morning state. However, if the coin lands heads, she gets strong evidence for *Heads & Tuesday* immediately after waking up on Tuesday. When she now wakes up on Monday, she ought at first to be indifferent between *Heads* and *Tails*; not getting the *Heads & Tuesday* evidence should then make her lean towards *Tails*.

Conservatives should agree. Both the update+revision rule and the principle of conservatism support thirding in the variant scenario. (Again, I spare you the details.) We can also grant that Beauty's evidence shortly after waking up on Monday is in all

15 Some authors, including [Kim 2009], [Schulz 2010], and [Spohn Forthcoming], have argued that general diachronic rules for self-locating beliefs support thirding. The problem with these arguments is that they assume diachronic norms can straightforwardly be applied to Sleeping Beauty. The result is thirding rather than the extreme form of halfing defended in [Hawley 2013] (on which Beauty should be certain that it is Monday) because the proposed update rules are incomplete: the new credences are not fully determined by the previous credences and the new evidence. One way of filling in the further parameter yields [Hawley 2013]'s position, another yields thirding.

relevant respects the same in the two versions of the story: in either case, the evidence rules out *Heads & Tuesday* and none of the other three possibilities. But conservatism denies that what Beauty should belief is fully determined by her evidence. In the original scenario, Beauty knew on Sunday that if the coin lands heads then her only epistemic successor will awaken on Monday. In the modified scenario, she is certain to also have an epistemic successor on Tuesday. For conservative agents, this difference in previous beliefs makes a difference to the later beliefs.[16]

A number of authors have argued that halfing would make Beauty vulnerable to diachronic Dutch Books – at least if she obeys "Causal Decision Theory" (see [Hitchcock 2004], [Draper and Pust 2008], [Briggs 2010]). The arguments have been contested (e.g. in [Bradley and Leitgeb 2006]), but I'm willing to grant that they work. From the conservative perspective I have outlined, they do not come as a surprise. After all, we have admitted that Beauty violates the ideal of diachronic rationality. If she were ideally rational, she would give credence 1/2 to *Heads*, be certain on Monday that it is Monday and on Tuesday that it is Tuesday. She would then not accept any of the Dutch Books that have been proposed. Since the scenario effectively stipulates that Beauty can't be ideally rational, we should not be surprised if that shows up in sub-optimal choices. It would still be a little surprising if thirding made Beauty invulnerable to Dutch Books, but that has never been shown, and as [Halpern 2006] points out, it is false: Beauty is also vulnerable to diachronic Dutch Books if she obeys thirding.[17]

Unlike the other cases discussed so far, the Sleeping Beauty problem is a scenario where information loss plays a central role. Let me therefore use the opportunity to make some general remarks on the alleged problem information loss poses for classical Bayesianism. Those who raise the problem often echo Timothy Williamson's claim that "forgetting is not irrational" [Williamson 2000: 219]. I disagree. I don't want to quibble over the word 'irrational'. Perhaps forgetting is not irrational in this or that sense of the term, but it is still a shortcoming of a broadly epistemic (rather than, say, prudential or

16 For analogous reasons, conservatism does not support halfing in [Bostrom 2007]'s variation *Beauty the High Roller*.

17 Halpern's presentation of his Dutch Book against thirding is terse and therefore often ignored or misunderstood. As I understand it, the setup goes as follows. On Sunday, Beauty is offered a deal in which she gains \$9 if the coin lands heads and loses \$8 if it lands tails. (She accepts the deal.) On every subsequent awakening, Beauty is offered to enter another deal that will be resolved on Wednesday. If she accepts on every awakening, the resolution is that she loses \$10 on heads and gains \$7 on tails. If she rejects the deal on at least one awakening, she neither gains nor loses anything. Now suppose that conditional on the hypothesis that the coin lands tails, Beauty is confident that she accepts the deal at the other – earlier or later – awakening. For thirders, accepting the deal at the current occasion then has a positive expected payoff of $1/3 \cdot \${-}8 + 2/3 \cdot \$7 = \$2$. That arguably supports the supposition we just made: since accepting has positive expected payoff, Beauty can be rationally confident that she accepts at any awakening. (This sort of feedback is discussed in [Skyrms 1990].) So she accepts. The bookie makes a sure profit of \$1.

moral) kind. We rightly criticise people if they forget important information.

The real problem with forgetting is that it is ubiquitous and unavoidable. Epistemically ideal agents never forget, yet we mortals can't help but do. If 'ought' implies 'can', it follows that conditionalization is not a norm for creatures like us. The same could be said for the first norm of classical Bayesianism, probabilistic coherence. Most people's degrees of belief fail to satisfy the laws of probability, and one might argue that full probabilistic coherence is an unattainable ideal.

If an ideal norm is out of reach, we have to think about fallback norms. The situation is familiar from ethics: if you can't bring yourself to be friendly, at least be polite; if you can't stop eating meat, at least don't buy from factory farms. A comprehensive moral theory should tell us not only what ideal agents would do, but also what one should do if one isn't ideal.

The problem of information loss therefore does point at a real problem for Bayesian epistemology. It does not, I think, show that there is anything wrong with the Bayesian picture of epistemically ideal agents. But it does remind us that a comprehensive epistemic theory should also give recommendations for agents who can't do what epistemically ideal agents would do.

So how should one deal with information loss? I'm afraid there is no simple, universal answer, if only because there are so many kinds of information loss. Opponents of classical Bayesianism tend to focus on artificial scenarios in which something that was once learned is completely forgotten, as if it had never been learned. A much more common situation is that something we once learned slowly begins to fade, even if some of its implications are retained. Our memory also tends to corrupt over time: we sometimes start believing things for which we never had any evidence. As a general principle, the right way to deal with these problems is to minimize the losses and errors. But that can mean different things, depending on the circumstances. The Sleeping Beauty problem provides a neat, although again quite idealized, illustration. It also illustrates that fallback norms are needed not only for cases where an agent actually can't live up to the ideal but also if there is merely a significant threat that living up to the ideal is impossible.

8 Evidentialism and time-slice epistemology

The principle of evidentialism is widely taken for granted in contemporary Bayesian epistemology. It is sometimes observed that the principle can clash with conditionalization, notably in cases where self-locating beliefs or information loss is in play, but that is considered a problem for conditionalization, not evidentialism (e.g. [Arntzenius 2003], [Moss 2015b]). I have presented a range of cases where evidentialism clashes not only with conditionalization, but with the basic conservative idea that the degree of belief one assigns to a proposition should not change in response to information that one previously

and rationally regarded as independent of that proposition – in particular, if one was rationally certain beforehand that one would receive the relevant information. In that sense, Fickle Frank receives no new information about theories A and B; Dr Evil learns nothing new about where he is; Fred receives no news about his machine; and Sleeping Beauty learns nothing new about the outcome of the coin toss. In each case, conservatism says that the agents should not change their relevant degrees of belief. Evidentialism (or permissive evidentialism, in the case of Fickle Frank) disagrees.

In scenarios like these, where evidentialism and conservatism come apart, conservative agents generally enjoy a practical advantage: it is easier for them to pursue long-term projects, and unless they deviate from ideal rationality in other respects they won't fall prey to Dutch Books. At least in the case of Dr Evil, we also saw that conservative agents may enjoy the epistemic advantage of having more accurate beliefs. In general, one can show that conservative norms such as conditionalization or the update+revision rule maximize the expected accuracy of the later beliefs as judged by the earlier beliefs (see [Greaves and Wallace 2006], [Leitgeb and Pettigrew 2010], [Schwarz 2015]). By contrast, self-aware evidentialist agents can find themselves in the curious position of planning to change their beliefs in a way that leads them further away from the truth.

The considerations of the present paper certainly don't refute the principle of evidentialism. But I hope they at least show that the principle is not an obvious triviality that can simply be taken for granted. It is not crazy to hold that one's beliefs should not always be proportioned to one's evidence.

To conclude, I want to briefly comment on a line of thought that is often taken to support evidentialism but that on reflection may actually support conservatism.

Evidentialists tend to focus on an agent at a particular time under particular circumstances and ask what the agent should believe, at that time, in those circumstances. Conditionalization seems to answer that the agent should set her new credences to equal her previous credences conditional on her present evidence. One then naturally worries how the agent (or her cognitive system) is supposed to obey this norm unless she happens to have access to her previous belief state. What if the agent does not have full information about her previous beliefs? In general, diachronic norms such as conditionalization seem to imply a strong kind of epistemic externalism according to which an agent's rational beliefs are directly constrained by her earlier beliefs, even if those are not recoverable from present evidence. One also worries how the external constraint is justified: what's so special about the agent's previous beliefs, as opposed to her later beliefs or other people's beliefs?

But conditionalization answers a different question. The question it answers is how an agent's belief state should change in order to accommodate new information. More generally, diachronic norms of rationality specify how an agent's attitudes should evolve from one time to another. In this dynamical process, the earlier belief state is not an

arbitrary and potentially inaccessible external factor: it is the starting point of the process.

Focusing on the dynamical question also puts the alleged problem of information loss in a different light. Conditionalization says that in order to accommodate new information, an agent's new credences should equal the old credences conditional on the new information. One may reasonably complain that this does not, say, take into account possible changes in the agent's location. But how could it be an objection that the agent then won't lose any information?

If we accept the dynamical question as legitimate – if we accept that there are epistemic norms on the dynamics of belief – evidentialism loses much of its appeal. It is therefore no surprise that philosophers who favour evidentialism have also declared the dynamical question illegitimate. According to what Sarah Moss ([2015b], [2015a]) calls "time-slice epistemology", whether an agent satisfies an epistemic norm is always a matter of the agent's state at a given moment in time, irrespective of her earlier or later state (see also [Hedden 2015b], [Hedden 2015a]). Time-slice epistemology rules out epistemic norms on the evolution of belief.

While I suspect that the time-slice perspective plays an important role in motivating evidentialism, the two positions are in principle independent. In fact, I would advise time-slice epistemologists to eschew evidentialism, and evidentialists to eschew time-slice epistemology. Let me explain.

Time-slice epistemology rules out genuine diachronic norms, but it allows for synchronic counterparts of diachronic norms. A synchronic counterpart of conditionalization, for example, is the principle of *reflection* ([van Fraassen 1984], [Goldstein 1983]) and its backwards-looking relative, the principle of *inverse reflection*:

$$Cr_t(A/Cr_{t-1}(A/E_t) = x) = x.$$

In words: if a rational agent's present evidence is E_t, then her credence in any proposition A, conditional on the assumption that her previous credence in A given E_t was x, should be x. Such principles obviously need qualification. It is easy to come up with scenarios where it would be irrational to defer to one's previous judgements – say, because one has reason to believe that these judgements were biased or caused by cognitive malfunctioning. We also need to make room for propositions that can change their truth-value. A more careful (informal) formulation of inverse reflection might therefore say that if an agent's present evidence is E_t, then her credence in any proposition A, conditional on the assumption that she previously *and rationally* assigned credence x to the hypothesis that *A was going to be the case* given that *E_t was going to be the case*, should be x.[18]

18 See [Schwarz 2012] for a more precise formulation. The need for the rationality qualification has often been noted in the context of the original reflection principle, see e.g. [Briggs 2009]; it is also well-known in discussions of conservatism in traditional epistemology, which often concentrate on

Whenever an agent has full information about her previous beliefs, satisfying inverse reflection is tantamount to obeying update+revision. Accordingly, inverse reflection sides with conservatism and against evidentialism in cases like *Fickle Frank*, *Dr. Evil*, and *The broken duplication machine*.[19] Since I believe that conservatism gets these cases right, I would suggest that even time-slice epistemologists should abandon evidentialism.

Conversely, I would advise evidentialists to not endorse time-slice epistemology. The reason turns on an important topic that I have largely tried to circumvent: the individuation of evidence. In the classical Bayesian picture, the evidence on which agents conditionalize is often taken as given by the agents' sensory experience. Such an experiential conception of evidence would be disastrous for evidentialism. Surely it is rational for me to assign high credence to the hypothesis that Luanda is the capital of Angola even at times where I have no relevant experience. Evidentialism requires a different conception of evidence. Perhaps our evidence is identified with our knowledge, or with everything of which we are certain, or perhaps is treated as an independent kind of propositional attitude. In each case, the evidence is not simply "given" from outside the epistemic domain. Thus one can allow for diachronic epistemic norms on the maintenance and update of evidence. For example, one may hold that if an agent knows that it is Sunday and that her next awakening will be on Monday, then (*ceteris paribus*) upon awakening she ought to know that it is Monday. In my view, it is highly plausible that there are such norms.

A full discussion of time-slice epistemology is beyond the scope of the present paper, but I want to emphasize that the hypothesis is deeply revisionary. Consider a collection of rational agents in different circumstances, with different beliefs, desires, memories, values, and plans. Now imagine a creature composed by stitching together one-second temporal segments of these agents, in random order. We would not recognize such a creature as an intentional agent, let alone a rational agent. Most of the psychological properties that make agents candidates of psychological and epistemic evaluation take time. Our everyday conception of a rational agent is a conception not of a time-slice but of a temporally extended agent. This begins on a very short timescale: we expect agents with certain perceptual experiences to *then* form corresponding beliefs. On a somewhat larger scale, we expect agents to utter coherent sentences that don't change grammatical structure and topic or end abruptly in the middle; we expect them to engage in other extended activities such as eating a meal, going to the railway station, or reflecting on difficult decisions. This gradually turns into our expectation that rational agents pursue long-term projects and goals. It is hard to see how these diachronic conditions on rational

synchronic principles; see e.g. [McCain 2008] on "defeat conditions".

19 The case of Sleeping Beauty is harder because Sleeping Beauty arguably does not have full information about her previous beliefs. In [Schwarz 2012] and [Schwarz 2015], I argue that inverse reflection is compatible with thirding.

agents could be derived from purely synchronic norms on time-slices, or why they should.

Absent strong reasons to the contrary, we should accept the dynamical question as legitimate. We should accept that are basic norms on the dynamics of belief. Where these norms clash with the principle of evidentialism, it is the latter that should be given up.

References

Frank Arntzenius [2003]: "Some problems for conditionalization and reflection". *Journal of Philosophy*, 100: 356–370

Nick Bostrom [2007]: "Sleeping Beauty and self-location: A hybrid model". *Synthese*, 157: 59–78

Darren Bradley [2011a]: "Confirmation in a Branching World: The Everett Interpretation and Sleeping Beauty". *British Journal for the Philosophy of Science*, 62: 323–342

— [2011b]: "Self-location is no problem for conditionalization". *Synthese*, 182: 393–411

Darren Bradley and Hannes Leitgeb [2006]: "When betting odds and credences come apart: More worries for Dutch book arguments". *Analysis*: 119–127

Richard Bradley [2005]: "Radical Probabilism and Bayesian Conditioning". *Philosophy of Science*, 72: 342–364

Rachael Briggs [2009]: "Distorted Reflection". *Philosophical Review*, 118(1): 59–85

— [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

Rudolf Carnap [1962]: "The Aim of Inductive Logic". In E. Nagel, P. Suppes and A. Tarski (Eds.) *Logic, Methodology and Philosophy of Science,* Stanford: Stanford University Press, 303–318

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

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

John Earman [1992]: *Bayes or Bust?*. Cambridge, MA: MIT Press

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

— [2010]: "Subjective Probabilities should be Sharp". *Philosophers' Imprint*, 10: 1–11

Richard Feldman and Earl Conee [1985]: "Evidentialism". *Philosophical Studies*, 48(1): 15–34

Michael Goldstein [1983]: "The Prevision of a Prevision". *Journal of the American Statistical Association*, 78: 817–819

Hilary Greaves [2007]: "On the Everettian epistemic problem". *Studies in History and Philosophy of Modern Physics*, 38: 120–152

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

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

James Hawthorne [2005]: "Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions". *Mind*, 114: 277–320

Brian Hedden [2015a]: *Reasons without persons: rationality, identity, and time.* Oxford University Press

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

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

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

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

Terry Horgan and Anna Mahtani [2013]: "Generalized Conditionalization and the Sleeping Beauty Problem". *Erkenntnis*, 78(2): 333–351

Colin Howson and Peter Urbach [1993]: *Scientific Reasoning.* La Salle (Ill.): Open Court Press, 2nd edition

David Hume [1777/1993]: *An Enquiry Concerning Human Understanding*. Indianapolis: Hackett Publishing Company

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

James Joyce [1999]: *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press

— [2005]: "How probabilities reflect evidence". *Philosophical Perspectives*, 19: 153–178

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

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

— [2010]: "Probability logic, logical probability, and inductive support". *Synthese*, 172(1): 97–118

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

— [1980]: "A Subjectivist's Guide to Objective Chance". In Richard Jeffrey (Ed.), *Studies in Inductive Logic and Probability* Vol. 2, University of California Press, Berkeley.

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

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

Patrick Maher [2010]: "Explication of Inductive Probability". *Journal of Philosophical Logic*, 29: 593–616

Kevin McCain [2008]: "The virtues of epistemic conservatism". *Synthese*, 164(2): 185–200

Christopher Meacham [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

— [2014]: "Impermissive Bayesianism". *Erkenntnis*, 79: 1185–1217

Sarah Moss [2015a]: "Credal dilemmas". Forthcoming in *Noûs*

— [2015b]: "Time-slice epistemology and action under indeterminacy". Forthcoming in *Oxford Studies in Epistemology*, volume 5

Mike Oaksford and Nick Chater [2007]: *Bayesian Rationality: The Probabilistic Approach To Human Reasoning*. Oxford: Oxford University Press

David Papineau and Víctor Durà-Vilà [2009]: "A thirder and an Everettian: a reply to Lewis's 'Quantum Sleeping Beauty"'. *Analysis*, 69: 78–86

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

Stuart J. Russell and Peter Norvig [2004]: *Artificial Intelligence: A Modern Approach*. Cambridge (MA): MIT Press, 2nd edition

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

Eric Schwitzgebel [2016]: "1% Skepticism". Forthcoming in *Noûs*

Brian Skyrms [1980]: "Higher Order Degrees of Belief". In D.H. Mellor (Ed.) *Prospects for Pragmatism,* Cambridge: Cambridge University Press

— [1987]: "Updating, Supposing, and Maxent". *Theory and Decision*, 22: 225–246

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

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

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

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

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

Michael G. Titelbaum [2013]: *Quitting Certainties*. Oxford: Oxford University Press

Hamid Vahid [2004]: "Varieties of epistemic conservatism". *Synthese*, 141(1): 97–122

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

David Wallace [2012]: *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford: Oxford University Press

Roger White [2005]: "Epistemic Permissiveness". *Philosophical Perspectives*, 19: 445–459

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

Jon Williamson [2011]: "Objective Bayesianism, Bayesian Conditionalisation and Voluntarism". *Synthese*, 178(1): 67–85

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