Sleeping Beauty and the Dynamics of *De Se* Beliefs

Christopher J. G. Meacham

Published in *Philosophical Studies*, 138 (2008): 245-269.

Abstract

This paper examines three accounts of the sleeping beauty case: an account proposed by Adam Elga, an account proposed by David Lewis, and a third account defended in this paper. It provides two reasons for preferring the third account. First, this account does a good job of capturing the temporal continuity of our beliefs, while the accounts favored by Elga and Lewis do not. Second, Elga's and Lewis' treatments of the sleeping beauty case lead to highly counterintuitive consequences. The proposed account also leads to counterintuitive consequences, but theyre not as bad as those of Elga's account, and no worse than those of Lewis' account.

1 Introduction

In standard possible worlds semantics, propositions are sets of possible worlds. To believe a proposition is to believe that your world is one of the worlds in that set. So the proposition that there are extraterrestrials is the set of worlds in which there are extraterrestrials, and to believe that there are extraterrestrials is to believe that your world is a member of that set.

A belief in a proposition is a belief about what the world is like. But in addition to beliefs about what the world is like, there are beliefs about where one is in the world. David Lewis (1979) has argued that these beliefs can't be expressed in terms of possible worlds. To accommodate beliefs about where we are in the world, Lewis proposed to extend standard possible worlds semantics by introducing *centered worlds*, possible worlds paired with individuals and times. A set of centered worlds is a *centered proposition*.[1] To believe a centered proposition is to believe that your current centered world is one of the centered worlds in that set. So the centered proposition that it's 9 am is the set of centered worlds at which it's 9 am, and to believe that it's 9 am is to believe that your current centered world is a member of that set.

Following Lewis, call beliefs that can be expressed in terms of possible worlds *de dicto beliefs*, and beliefs that can be expressed in terms of centered worlds *de se beliefs*. In his paper Lewis raises the question of what happens to Bayesian decision theory when we consider *de se* beliefs instead of *de dicto* beliefs. His answer is a natural one:

[1] Lewis himself calls them *properties*.

> "Very little. We replace the space of worlds by the space of centered worlds, or by the space of all inhabitants of worlds. All else is just as before."[2]

However, this answer is untenable. When you update your beliefs using standard Bayesian conditionalization your certainties are permanent: if you're certain a proposition is true before updating then you'll be certain it's true after updating. So on the account Lewis suggests, if you're certain that a centered proposition is true you will always remain certain that it's true. But suppose you're looking at a clock you know is accurate. If the clock reads 9 am, then you're certain of the centered proposition that it's 9 am. Given Lewis' suggestion, since you're certain of the centered proposition that it's 9 am when the clock reads 9 am, you should always remain certain that it's 9 am. So you should remain certain that it's 9 am a minute later, when the clock reads 9:01 am. Obviously, this is not how our beliefs should be updated.[3] We need a more sophisticated dynamics for *de se* beliefs.

Lewis (2001) himself employs a more sophisticated dynamics in his discussion of the sleeping beauty case. This case raises precisely the issue of how *de se* beliefs should change over time. By looking at the different treatments of the case, we can gain insight into the dynamics of *de se* beliefs. In this paper I'll look at three accounts of the sleeping beauty case: an account proposed by Adam Elga (2000), an account proposed by David Lewis (2001), and a third account I'll defend in this paper.

I'll offer two reasons for preferring my account over theirs. First, every account of what our credences should be must accommodate the temporal continuity of our credences. I'll show that the dynamics I propose do a better job of capturing our intuitions about how our credences should be diachronically coordinated than the dynamics favored by Elga and Lewis. Second, I'll argue that Elga's and Lewis' treatments of the sleeping beauty case lead to highly counterintuitive consequences. I'll show that the account I offer also leads to counterintuitive consequences, but I'll argue that they're not as bad as those of Elga's account, and no worse than those of Lewis' account.

There are several other considerations that can be used to assess the merits of these accounts, such as betting arguments, considerations regarding reflection, etc. These issues are important, and have been addressed in a number of places.[4] They are beyond the limited scope of this paper, however, and I won't look at them here.

The rest of this paper will proceed as follows. In the next section I'll look at some natural dynamics for *de se* beliefs. In the third section I'll discuss the temporal continuity of beliefs, and I'll show that the dynamics I propose can account for much of this continuity, while the dynamics of Elga and Lewis cannot. In the fourth section I'll present the sleeping beauty case and sketch the three responses

[2] Lewis (1979), p. 149.

[3] Arntzenius (2003), Halpern (2004) and Hitchcock (2004) have noted this problem with extending standard conditionalization to *de se* beliefs.

[4] For a sampling of this literature, see Elga (2000), Lewis (2001), Arntzenius (2002), Dorr (2002), Arntzenius (2003), Halpern (2004) and Hitchcock (2004). The article by Halpern is especially relevant to this paper, as he defends an account similar to the account defended here. As a result, much of what he says, such as his treatment of betting arguments, reflection, etc., applies to my account as well.

to it. In the fifth and sixth sections I'll look at Elga's and Lewis' responses in detail, and show how they both lead to counterintuitive consequences. In the seventh section I'll critically examine my account by looking at some consequences of it that are also counterintuitive. In the eighth section I'll sum up my conclusions.

2 Belief Dynamics

It's standard to assume that belief is not an all-or-nothing affair, but rather admits of degrees. A subject's beliefs are represented by a probability function over the space of possibilities. The function assigns values between zero and one to regions of the space, representing the subject's confidence that some possibility in that region obtains. The values it assigns are countably additive: the value of the union of countably many non-overlapping regions of the space is the sum of the values of each of these regions. The value it assigns to the entire space of possibilities is one, representing the subject's certainty that some possibility or other obtains.

In the case of *de dicto* beliefs, the space of possibilities is the space of possible worlds. The belief function takes worlds as arguments, and assigns to each world a degree of belief, or credence. The credence assigned to a proposition is the sum of the credences assigned to the worlds in that proposition.[5] The worlds in which the subject has non-zero credences are the worlds she thinks might be hers, or her *doxastic worlds*.[6]

When we generalize to *de se* beliefs, the space of possibilities becomes the space of centered worlds. The belief function takes centered worlds as arguments, and assigns to each centered world a credence. The credence assigned to a centered proposition is the sum of the credences assigned to the centered worlds in that centered proposition. The centered worlds in which the subject has non-zero credences are the centered worlds she thinks might be hers, or her *doxastic alternatives*.

Let's look at the dynamics of *de dicto* belief. At the core of an account of belief is an updating rule, a rule for generating new credences. The canonical updating rule is conditionalization. For simplicity let's focus on standard conditionalization, ignoring Jeffrey conditionalization and the like.[7]

On standard conditionalization you generate your current credences from your prior credences and your current evidence. To get your new credences you take your prior credences, set the credence in every world incompatible with your evidence to 0, and then normalize the credences in the remaining doxastic worlds; i.e., adjust the values such that they sum to 1, and such that the ratios between them are the same as the ratios between your prior credences. This way of updating makes certainties permanent. This is because you can only lose doxastic worlds

[5]Throughout the paper I'll ignore the complications that arise when we consider uncountably infinite numbers of worlds, and which we need measure theory to properly address.

[6]Having a non-zero credence in a world and thinking a world might be yours are only equivalent if, like Lewis, one adopts a non-standard account of credence. Those inclined to reject this view can just take a subject's doxastic worlds to be the worlds in which she has a non-zero credence. (Similar comments obtain for my description of doxastic alternatives, below.)

[7]See Howson and Urbach (1993) for a description of Jeffrey conditionalization.

in this process, not gain them. Being certain of a proposition P entails that all of your current doxastic worlds are compatible with P, and if you only lose doxastic worlds when you update, then all of your future doxastic worlds will be compatible with P as well.

There is another version of *de dicto* conditionalization that does not have this consequence. Call it *hp-conditionalization*. On hp-conditionalization, the role of prior credences in standard conditionalization is played by a hypothetical initial credence function. The notion of a hypothetical initial credence function is not new; hypothetical initial credence functions of various kinds have been employed in a number of places in the Bayesian literature.[8] In this context I'll take a subject's hypothetical initial credence function, or *hypothetical priors*, to be normative values: they are the credences she ought to have if she had no evidence whatsoever. So, like her credences, they are a set of values assigned to a space of possibilities that satisfies the probability axioms. But, unlike her credences, these values are static.

On hp-conditionalization you generate your current credences from your hypothetical priors and your current evidence. To get your new credences you take your hypothetical priors, set the credence in every world incompatible with your evidence to 0, and then normalize the credences in the remaining doxastic worlds; i.e., adjust the values such that they sum to 1, and such that the ratios between them are the same as the ratios between your hypothetical priors.

Unlike standard conditionalization, hp-conditionalization does not make certainties permanent, since subjects can lose *and* gain doxastic worlds. If your current evidence is compatible with worlds that your previous evidence was not, then you gain doxastic worlds when you update. This can happen, for example, when a subject suffers memory loss. If you have a perfect memory then your current evidence will include your memory of the previous evidence you've received, and you'll usually continue to rule out possibilities that you've ruled out in the past.[9] But if you forget evidence that ruled out certain worlds, then your current evidence will no longer rule those worlds out.[10]

How should we generalize conditionalization to *de se* beliefs? As we saw in section 1, we cannot simply start with standard conditionalization and replace the space of worlds with the space of centered worlds. In the context of *de se* beliefs we both gain and lose possibilities, but standard conditionalization only allows the loss of possibilities.

We get a more promising account if we start with hp-conditionalization and replace worlds with centered worlds. Call this version of *de se* conditionalization

[8]For example, see Earman (1992) and Howson and Urbach (1993) on the problem of old evidence, Strevens (2004) on inductive frameworks, Maher (2005) on the confirmation relation, Lewis (1980) and Hall (1994) on the relation between credence and chance, and Bartha and Hitchcock (1998) on the Doomsday Argument. Worries about the use of hypothetical initial credence functions of various kinds, and the pros and cons of adopting them, are discussed extensively in these references.

[9]There can be exceptions to this if there are centered worlds that (i) you have a non-zero prior in, (ii) are subjectively identical to your current state of perfectly remembering your past evidence, and (iii) are located at worlds which you had previously eliminated.

[10]In a similar fashion, hp-conditionalization can deal with cases like the Shangri-la case given by Arntzenius (2003).

centered conditionalization. On centered conditionalization you generate your current credences from your hypothetical priors and your current evidence. To get your new credences you take your hypothetical priors in centered worlds, set the credence in every centered world incompatible with your evidence to 0, and then normalize the credences in the remaining doxastic alternatives; i.e., adjust the values such that they sum to 1, and such that the ratios between them are the same as the ratios between your hypothetical priors.

Centered conditionalization is one way to modify hp-conditionalization in order to account for *de se* beliefs. However, centered conditionalization and unmodified hp-conditionalization are incompatible. To see this, consider a subject with just two doxastic worlds, A and B, with two doxastic alternatives at each world. Assume that her credences are divided equally among alternatives, so that her credence in each alternative is $\frac{1}{4}$ and her credence in each world is $\frac{1}{2}$. What should her credences in worlds A and B be if one of her alternatives at A is eliminated? According to hp-conditionalization her credences in A and B should remain $\frac{1}{2}/\frac{1}{2}$. Her evidence hasn't eliminated any doxastic worlds, so hp-conditionalization will assign the same credences. According to centered conditionalization, on the other hand, her credences in A and B should change. After the alternative at A is eliminated, centered conditionalization redistributes this credence among alternatives, so that her credence in each alternative is $\frac{1}{3}$. Since she has one alternative at A and two alternatives at B, her credence in A should now be $\frac{1}{3}$ and her credence in B should now be $\frac{2}{3}$.

There is another way to modify hp-conditionalization in order to accommodate *de se* beliefs that avoids this conflict. I'll call it *compartmentalized conditionalization*. On compartmentalized conditionalization you use hp-conditionalization to assign your credence in worlds, and then you divide your credence in each world among its centered worlds in proportion to your centered world's priors. So on compartmentalized conditionalization, your credences are determined by your priors and your current evidence. (Recall that your priors, like any probability function, are additive, so your prior in a world is the sum of your priors in the centered worlds at that world.)

Here's another way to describe compartmentalized conditionalization. Given your priors and current evidence, compartmentalized conditionalization tells you to determine your new credences in three steps. First, you take your hypothetical priors, and set the credence in every centered world incompatible with your current evidence to 0. Second, you normalize the credences in the remaining doxastic worlds; i.e., adjust the values assigned to each doxastic world such that they sum to 1, and such that the ratios between them are the same as the ratios between their priors. Finally, you normalize your credences in the remaining doxastic alternatives at each world; i.e., at each world adjust the values assigned to the alternatives so that they sum to the credence assigned to that world, and such that the ratios between them are the same as the ratios between their priors.[11]

[11] In an intermediate draft of this paper, in an attempt to make things easier to follow, I replaced this rule with a simpler rule which divides the credence assigned to a world equally among the remaining doxastic alternatives at that world, instead of in proportion to their priors. But this rule is incompatible with standard characterizations of hypothetical priors. I.e., if *hp* is your priors and you have no evidence, then this rule will not yield *hp* as your credences again.

We'll see some examples of how compartmentalized conditionalization works in sections 3 and 4.

In this paper I'll look at three accounts of the sleeping beauty case: my account, Elga's account and Lewis' account. An account of sleeping beauty requires an updating rule for *de se* beliefs, as well as further constraints on a subject's credences. The account I'm defending in this paper employs compartmentalized conditionalization and the Principal Principle. (Someone who adopts compartmentalized conditionalization may also want to adopt what I'll call a 'Continuity Principle', but this principle isn't needed for the sleeping beauty case.) Elga's and Lewis' accounts both employ centered conditionalization as their updating rule, but differ on the other principles they adopt. We'll see what additional principles Elga's and Lewis' accounts employ in sections 5 and 6.

3 Continuity

3.1 Continuity and the Passage of Time

De se beliefs raise questions about belief continuity which don't arise in *de dicto* contexts. Consider again the case presented in the introduction, where a subject is watching a clock she knows to be accurate. When the clock changes from 9 am to 9:01 am, the subject discards all of her alternatives at which it's 9 am and replaces them with alternatives at which it's 9:01 am. It seems that her credence in these new alternatives should bear some relation to her credence in the alternatives they've just replaced. But nothing we've said so far requires that this be the case.

Suppose, for example, that the subject watching the clock has only two doxastic worlds, A and B, and that she has only one doxastic alternative at each world. Further suppose that she updates her beliefs using centered conditionalization, and that at 9 am her priors in her two alternatives (A(9:00) and B(9:00)) are equal, so her credences in A(9:00) and B(9:00) are $\frac{1}{2}/\frac{1}{2}$. When she sees the clock register 9:01 am, what should her credences in A(9:01) and B(9:01) be? Intuitively, they should be $\frac{1}{2}/\frac{1}{2}$. But there is no reason they have to be this way. Although her priors in A(9:00) and B(9:00) are equal, at 9:01 am these are no longer her alternatives. Her alternatives are now B(9:01) and B(9:01), and nothing we've said so far forces her to have equal priors in these alternatives.

For subjects like us, who have a sense of time passing, every belief change will include a time changing component. As we notice time pass, we replace our old alternatives with new ones located at a later time. Since every evidential change brings an awareness that time has passed, every belief change involves the replacement of old alternatives with new ones. Nothing we've said so far entails that the beliefs of such subjects will be in any way constant—that their credences won't fluctuate wildly simply due to the passage of time. But we intuitively think that there should be such constraints; constraints which require a rational subject's beliefs to be diachronically coordinated in the appropriate way. Call constraints of this kind *Continuity Principles*.

A Continuity Principle will take the following form: a subject's credences in her alternatives before and after a belief change ought to be diachronically coordinated when those alternatives are suitably related. For convenience, let us say that

an old and new alternative which are suitably related are *continuous* with one another. To obtain a specific Continuity Principle we need to answer two questions. First, under what conditions are a pair of alternatives continuous? Second, given that a pair of alternatives are continuous, how should our credences in them be correlated?

Let's start with the first question: under what conditions are a pair of alternatives continuous? A necessary condition for continuity is that the alternatives must be located at the same world; we want A(9:00) to be continuous with A(9:01), not B(9:01). A trivial sufficient condition for continuity is identity; if a centered world survives a belief change, then it is clearly continuous with itself.[12]

But these two conditions leave a number of difficult cases undecided. Consider a subject about to undergo duplication. Before duplication she has one doxastic alternative at each of her doxastic worlds. After duplication she'll have two alternatives, one centered on the original individual and one centered on the duplicate. How should her credences in these two new alternatives be related to her credence in her original alternative? When we consider the original individual, it seems that the new alternative should be continuous with the old alternative she had before duplication, since it's just its temporal successor. But it's less clear what to think when we consider the duplicate individual. Should this new alternative be continuous with the old alternative of the original individual? I think it's not obvious what to say.

There are a number of other hard cases to consider, such as cases of fission, fusion, the addition and elimination of alternatives located at different times, and so on. These cases make it difficult to spell out complete necessary and sufficient conditions for the continuity of alternatives. Other than the necessary condition given above, I won't take a stand in this paper on what the criteria for continuity should be. Instead, I'll allow for a variety of Continuity Principles, differing in what standard of continuity they employ. When we come to an argument that requires a Continuity Principle of some kind, I'll provide explicit standards of continuity that are sufficient for these arguments to go through.

Let's turn to the second question: given that a pair of alternatives are continuous, how should their credences be related? Consider again the case of the subject watching a clock. In this case we're naturally inclined to assume that her A(9:00) and B(9:00) alternatives are continuous with her A(9:01) and B(9:01) alternatives, respectively. Intuitively, what does this entail about her credences in these alternatives? It seems as if her credences in the new alternatives should be the same as her credence in the earlier alternatives they're continuous with. So if her credences in A(9:00) and B(9:00) are $\frac{1}{2}/\frac{1}{2}$, her credences in A(9:01) and B(9:01) should be $\frac{1}{2}/\frac{1}{2}$ as well.

Of course, we don't want to require that credences in continuous alternatives always be the same. Suppose that at 9:01 am the subject learns ¬B, and so has only one alternative at 9:01 am, A(9:01). A(9:01) is continuous with A(9:00), but her credence in A(9:01) should be 1, not $\frac{1}{2}$. So we don't want continuous

[12] On a natural picture of evidence the only belief changes on which old and new alternatives pick out the same centered world are the trivial belief changes that leave one with the same alternatives. (See the following footnote.) It's not clear that such belief changes are possible for subjects who have a sense of time passing.

alternatives to always be assigned the same credences, just to be assigned the same credences when they're in similar evidential situations.

We can capture this intuition by requiring continuous alternatives to have the same priors. Both centered and compartmentalized conditionalization are hypothetical prior rules; given one's evidence, they assign credences to alternatives in a manner determined by their priors. So we can get continuous alternatives to have appropriately coordinated credences by requiring them to have the same priors.

But this turns out to be a stronger constraint on priors than we need. This is because we can get the same constraint on credences with a strictly weaker constraint on priors. Let's see how to do this for the two rules we're concerned with.

On centered conditionalization, a subject's credences are distributed among her doxastic alternatives in proportion to her priors in those alternatives. Thus the amount of credence assigned to an alternative isn't sensitive to the absolute magnitude of the alternative's prior, only to the *ratio* between its prior and the priors of the other alternatives. So all we need to keep track of is the ratios of the priors between alternatives. Thus on centered conditionalization, the Continuity Principle just requires that the ratio of priors between new alternatives be the same as the ratio of priors between any old alternatives they're continuous with.[13]

To see this, consider again the case of a subject watching a clock. Let her prior in A(9:00) and B(9:00) be x. Since centered conditionalization assigns credences to alternatives in proportion to their priors, her credence at 9 am in A(9:00) and B(9:00) will be $\frac{1}{2}/\frac{1}{2}$. If her prior in A(9:01) and B(9:01) is also x, then her 9:01 am credences in A(9:01) and B(9:01) will also be $\frac{1}{2}/\frac{1}{2}$, as desired. But if her prior in both A(9:01) and B(9:01) was $2x$, her 9:01 am credences in A(9:01) and B(9:01) would still be $\frac{1}{2}/\frac{1}{2}$, since the ratio between their priors is the same. So to get continuity, we just need the ratio of priors between new alternatives be the same as the ratio of priors between the old alternatives they're continuous with.

On compartmentalized conditionalization, a subject's credences are distributed among worlds in proportion to her priors in those worlds, and her credence at each world is divided among its alternatives in proportion to her priors in those alternatives. Thus the proportion of a world's credence assigned to an alternative

[13]It may be helpful to see explicitly how one would go about imposing this constraint on priors. Doing this requires some decisions about one's background assumptions, and I'll work with the following simple model: a subject's evidence is their subjective state, so the centered worlds compatible with a subject's evidence are those centered on individuals in the subject's subjective state. Let the *subjective state sets* S_i be sets containing all of the centered worlds centered on a possible individual in a given subjective state.

To check whether your priors satisfy the constraint, do the following, for every possible belief change. Let the subjective state of the subject before the belief change be S_i, and the subjective state of the subject after the belief change be S_j. Use your criteria for continuity to determine whether any of the centered worlds in S_i (the alternatives compatible with your evidence before the belief change) would be continuous with any of the centered worlds in S_j (the alternatives compatible with your evidence after the belief change), if you were to undergo such a belief change. Then take the centered worlds in S_i that are potentially continuous with centered worlds in S_j, and make sure that the priors assigned to these S_i centered worlds are such that the ratios between them are the same as the ratios between their S_j successors.

isn't sensitive to the absolute magnitude of the alternative's prior, only to the ratio between its prior and the priors of the other alternatives at that world. So all we need to keep track of is the ratios of the priors between alternatives at each world.[14] Thus on compartmentalized conditionalization, the Continuity Principle just requires that the ratio of priors between the new alternatives at each world be the same as the ratio of priors between any old alternatives at that world they're continuous with.

To see this, consider again the case of a subject watching a clock, but this time let her have two alternatives at each world, A(9:00) and A′(9:00) at A, and B(9:00) and B′(9:00) at B. Let her prior in worlds A and B be y, and her prior in each of these four centered worlds be x. Since compartmentalized conditionalization assigns credences to worlds in proportion to their priors, her credence in A and B will be $\frac{1}{2}/\frac{1}{2}$ at both at 9 am and 9:01 am. Since compartmentalized conditionalization divides the credence of a world among its alternatives in proportion to their priors, her 9 am credence in each world will be split evenly between the two alternatives at that world, and her 9 am credence in each alternative will be $\frac{1}{4}$. Now, if her prior in each of the four temporal successors to these alternatives (A(9:01), A′(9:01), B(9:01) and B′(9:01)) is also x, then her 9:01 am credence in these successors will also be $\frac{1}{4}$, as desired. But if her prior in A(9:00) and A′(9:00) was $\frac{1}{2}x$, and her prior in B(9:00) and B′(9:00) was $2x$, her 9:01 am credences in these successors would still be $\frac{1}{4}$. Her credence in A and B will be $\frac{1}{2}$, and this will be divided evenly between the two alternatives at each world. So to get continuity, we just need the ratio of priors between new alternatives at each world to be the same as the ratio of priors between the old alternatives they're continuous with.

Notice that compartmentalized conditionalization requires a strictly weaker constraint on priors than centered conditionalization in order to satisfy the Continuity Principle. Compartmentalized conditionalization is better at capturing our intuitions about how our credences should be diachronically coordinated, and so it requires fewer constraints on priors to keep our credences in line. In the next section we'll see why this is so, and we'll take a careful look at the extent to which compartmentalized conditionalization does a better job of capturing these intuitions.

3.2 Continuity and Dynamics

To what extent do centered and compartmentalized conditionalization need a Continuity Principle in order to capture our intuitions about how our credences should evolve? The former badly needs a Continuity Principle in order to get acceptable credal behavior; without it our credences can vary arbitrarily without constraint. The latter, on the other hand, does well without a Continuity Principle. On compartmentalized conditionalization our credences in worlds aren't subject to arbitrary variation, and this limits the potential for arbitrary variation in our credences in alternatives. For subjects like us, this results in naturally coordinated credences for almost all of our alternatives. Let's look at these claims in more detail.

[14] What about the ratios of priors between worlds? We don't need to put constraints on these because worlds don't get replaced by temporal successors, so the ratios between their priors are static.

Unlike centered conditionalization, compartmentalized conditionalization naturally coordinates our credences in worlds. On centered conditionalization our credence in a doxastic world hangs on the priors of our current alternatives at that world, so as our alternatives change, our credence in the world can fluxuate wildly. On compartmentalized conditionalization our credence in a doxastic world hangs on the prior of that world, and as this value is static, our credence isn't subject to arbitrary variation.

Let's look at an example of how compartmentalized conditionalization coordinates our credences in worlds, and centered conditionalization does not. Consider again the subject who is watching a clock she knows to be accurate, and who has two doxastic worlds, A and B. At 9 am she has one doxastic alternative at each world, A(9:00) and B(9:00), and has equal credence in each. When she sees the clock register 9:01 am she'll replace each of her 9 am alternatives with a 9:01 am alternative. What do centered and compartmentalized conditionalization require of her 9:01 am credences?

If she's a centered conditionalizer, the fact that her 9 am credences in A(9:00) and B(9:00) were equal entails that her priors in A(9:00) and B(9:00) must be equal. But this doesn't say anything about her priors in A(9:01) or B(9:01). So if she's a centered conditionalizer her credences in the A and B worlds at 9:01 am can be completely unrelated to her credences in A and B at 9 am.

If she's a compartmentalized conditionalizer, the fact that her 9 am credences in A(9:00) and B(9:00) are equal entails that her priors in the worlds A and B are equal, although her priors in the centered worlds A(9:00) and B(9:00) may not be. This doesn't say anything interesting about her priors in A(9:01) or B(9:01), of course, but it doesn't matter.[15] Her credence in A and B at 9:01 am will be $\frac{1}{2}/\frac{1}{2}$ regardless of her priors in A(9:01) and B(9:01). So if she's a compartmentalized conditionalizer she'll naturally have coordinated credences in A and B.

This coordination of our credences in worlds substantially restricts the potential for arbitrary variation in our credences in alternatives. To see this, let's look at what kinds of arbitrary variation compartmentalized conditionalization allows. At doxastic worlds with a single alternative, the alternative is assigned all of the world's credence. Since a lone alternative and its temporal successor will both be assigned the full credence of the world, and the credences of worlds are intuitively coordinated, the pair of alternatives will have intuitively coordinated credences as well. So there won't be arbitrary variation in the credence of alternatives at single alternative worlds. The only place where arbitrary variation can arise is at doxastic worlds with multiple alternatives. At multiple alternative worlds the credence of a world is divided among alternatives in proportion to their priors. If the priors of temporal successors have different relative magnitudes than their predecessors, they'll be assigned different proportions of the world's credence, and the credences of the old and new alternatives won't be intuitively coordinated.

So on compartmentalized conditionalization, arbitrary variations in the credences of alternatives can only happen at multiple alternative worlds. And unlike centered conditionalization, the amount of arbitrary variation is restricted to how

[15]Her prior in A and B does tell us some *uninteresting* things about her priors in A(9:01) and B(9:01), of course. Since the priors of worlds are equal to the sum of the priors of the centered worlds at that world, we know that $\text{hp}(\text{A}) \geq \text{hp}(\text{A(9:00)}) + \text{hp}(\text{A(9:01)})$, for example.

the credences of worlds are divided among their alternatives.

Consider again a case where a subject is looking at a clock they know to be accurate. As before, let her have two doxastic worlds at 9 am, A and B. This time, however, let her have one alternative at A and two alternatives at B: one centered on her, and one centered on a duplicate of her. At 9 am let her credence in A and B be $\frac{1}{2}/\frac{1}{2}$, and her credence in the two alternatives at B be $\frac{1}{4}/\frac{1}{4}$. Now, at 9:01 am the clock changes, and she replaces each of her 9 am alternatives with a 9:01 am alternative. What should her 9:01 am credences be like according to centered and compartmentalized conditionalization?

If she's a centered conditionalizer, the ratio of her credences in her alternatives at 9 am will be the same as the ratio of her priors in those alternatives. So the ratio of her priors in A(9:00), B(9:00) and B′(9:00) will be 2:1:1. This doesn't entail anything about her priors in their 9:01 am successors, however, and her credence at 9:01 am in each of the 9:01 am alternatives might be anything between 0 and 1.

If she's a compartmentalized conditionalizer, having equal credence in A and B at 9 am at entails that her priors in A and B are the same. Likewise, having equal credence in B(9:00) and B′(9:00) at 9 am entails that her priors in B(9:00) and B′(9:00) are the same. Her 9 am credences don't tell us anything about her prior in A(9:00), however, since her credence in A(9:00) will just be her credence in A regardless. As with centered conditionalization, none of this tells us anything interesting about her priors in her 9:01 am alternatives. But her priors in the A and B worlds will be the same, so her credence at 9:01 am in A and B will be the same as well: $\frac{1}{2}/\frac{1}{2}$. The stability of her credence in worlds imposes stability on her credences in alternatives. At single alternative worlds like A, there is no potential for arbitrary variation: the alternative at that world, A(9:01), will just be assigned the credence of that world, $\frac{1}{2}$. At multiple alternative worlds like B, there is potential for arbitrary variation. If B(9:01) and B′(9:01) have different priors, they'll be assigned different proportions of B's credence. But this isn't the extreme variation allowed by centered conditionalization; it's not the case that her credence in each alternative might be anything between 0 and 1. The only variation compartmentalized conditionalization allows is in how the credence of a world is divided among the alternatives at that world. In this case, her 9:01 am credences in B(9:01) and B′(9:01) are restricted to values between 0 and 1/2.

For subjects like us, the natural constraints on arbitrary variation imposed by compartmentalized conditionalization lead to almost complete credence coordination. There is only potential for arbitrary variation at doxastic worlds with multiple alternatives, and for subjects like us, such worlds are rare.

One way to see how rare doxastic worlds with multiple alterantives are is to note that many cases which may seem to involve multiple alternatives do not. Consider the following case. As I'm writing this, I'm wondering what time it is. When I last looked at the clock it was 6 pm, but I'm now unsure as to whether it's 7 pm or 7:05 pm. It might seem like this is a case where I now have two alternatives at each of my doxastic worlds; one located at 7 pm and another located at 7:05 pm. But there is a fact about the temporal distance between when I last looked at the clock and when I typed the sentence "As I'm writing this, I'm wondering what time it is." The doxastic alternatives where it's 7 pm are at doxastic worlds

where an hour has passed between these two events, while the doxastic alternatives where it's 7:05 pm are at doxastic worlds where 65 minutes have passed between these two events. So these two alternatives aren't at the same doxastic world after all, they're at different doxastic worlds.

Here's another way to see how rare multiple alternative worlds are. Your doxastic alternatives are the centered worlds that you think might be yours. Assume transparency—that subjects always have access to their own subjective states. Then, if you to think that a centered world is yours, it must be subjectively indistinguishable from your current subjective state. So all of the centered worlds you think might be yours—your doxastic alternatives—must be subjectively indistinguishable.

Worlds at which I currently have multiple doxastic alternatives are strange worlds; they are worlds at which there are multiple subjective states indistinguishable from my subjective state. Indeed, worlds at which I *ever* have multiple alternatives are strange worlds. Consider my life as a sequence of time-slices. Ignore times when I've been unconscious or otherwise incapable of rational thought, and consider slices that are far enough apart to be noticeably distinct. How many of these me-slices are in subjectively indistinguishable states? If I'm in the set of worlds I think I'm probably in, none of them are. Likewise, if the world is like I think it probably is, no me-slice will be in a state indistinguishable from that of any time slice of anyone else, present, future or past.

Without the addition of a Continuity Principle, centered conditionalization does nothing to keep our credences coordinated in an intuitive manner; it allows our credences to vary arbitrarily without constraint. Compartmentalized conditionalization, on the other hand, does a great deal to keep our credences coordinated. If we adopt compartmentalized conditionalization, then for the majority of our doxastic worlds—worlds at which we have a single alternative—the diachronic coordination of our credences falls right out of the dynamics. And at the rest of our doxastic worlds—strange worlds with multiple alternatives—the potential for arbitrary variation of our credences is severely constrained.

What bearing should this have on our evaluation of accounts that adopt centered and compartmentalized conditionalization? In general we evaluate philosophical proposals according to how well they fit two criteria. First, we evaluate how well the account fits our intuitions. Second, we evaluate the theory in terms of something like simplicity and elegance: we prefer simple theories over complicated ones, natural theories over contrived ones, and systematic theories over gerrymandered ones. Often we can modify a theory in order to alleviate deficits in one criterion at the cost of deficits for the other. We can add specially tailored principles to make an account more intuitive at the cost of making it more complicated and contrived, and we can streamline a theory by abandoning clauses and caveats at the cost of making it less intuitive.

Restricting ourselves to intuitions regarding continuity, let's apply these criteria to centered and compartmentalized conditionalization. Considered by themselves, both are elegant theories that do well with the second criterion, but centered conditionalization does worse with the first criterion: it does a poor job of capturing our intuitions about how our credences should be diachronically coordinated. If we add specially tailored Continuity Principles to these theories, then both do

well with the first criterion, but the centered conditionalization-and-Continuity Principle package does worse with the second: both theories impose a contrived constraint on priors, but with the centered conditionalization package the constraint on priors is much greater. Either way, continuity considerations give us a reason to favor an account that adopts compartmentalized conditionalization over an account that adopts centered conditionalization. And in the last section of this paper, we'll be in a position to see that compartmentalized conditionalization has yet further advantages regarding continuity.

4 Sleeping Beauty

An interesting case of *de se* belief change is the sleeping beauty case:

> *The Sleeping Beauty Case:* Some researchers are going to put you to sleep for several days. They will put you to sleep on Sunday night, and then flip a coin. If heads comes up they will wake you up on Monday morning. If tails comes up they will wake you up on Monday morning and Tuesday morning, and in-between Monday and Tuesday, while your are sleeping, they will erase the memories of your waking.

When you wake up there is no way for you to know if it is Monday or Tuesday. If you are in the world in which the coin came up heads, then it's Monday. If you are in the world in which the coin came up tails, then it may be Monday or Tuesday. Suppose you then learn that it's Monday. Then you'll know what day it is, but you still won't know whether the coin came up heads or tails. There are two questions to ask here. First, what should your credences be when you wake up? Second, what should your credences be if you learn that it's Monday?

Let's look at what my account says. Assume the Principal Principle, that a subject's credences should line up with what she thinks the chances are (if she's not in possession of inadmissible information).[16] On compartmentalized conditionalization a subject first divides her credences among worlds, and then divides the credence of each world equally among the alternatives at that world. So a subject's credence in worlds, and thus in *de dicto* propositions, only changes when she gains or loses doxastic worlds.

On Sunday you will have a $\frac{1}{2}/\frac{1}{2}$ credence that the coin toss came up heads/tails by the Principal Principle, with one doxastic alternative at each of your doxastic worlds. When you wake up on Monday you have one alternative (Monday) at each heads world and two alternatives (Monday and Tuesday) at each tails world. But although your doxastic alternatives have changed, you have the same doxastic worlds you had on Sunday. Since your doxastic worlds have remained the same, you will have the same credence in heads/tails: $\frac{1}{2}/\frac{1}{2}$. How should your $\frac{1}{2}$ credence in tails be divided between Monday and Tuesday? On compartmentalized conditionalization your credence in tails is divided between these two alternatives in proportion to their priors. So if their priors are (say) equal, your credence in Monday/Tuesday given tails will be $\frac{1}{4}/\frac{1}{4}$.

[16] See Lewis (1980).

What if you then learn that it's Monday? This eliminates the Tuesday alternative at your tails worlds, but doesn't eliminate any doxastic worlds. So again, your credence in heads/tails will remain the same: $\frac{1}{2}/\frac{1}{2}$.

What do Elga and Lewis say about the sleeping beauty case? Elga (2000) proposes that upon waking we should have a $\frac{1}{3}$ credence in heads and a $\frac{2}{3}$ credence in tails, the latter split evenly between Monday and Tuesday. If you learn that it's Monday, you should be a centered conditionalizer and regain your original $\frac{1}{2}/\frac{1}{2}$ credence in heads/tails.

Lewis (2001) proposes that we retain our $\frac{1}{2}/\frac{1}{2}$ credence in heads/tails when we wake up, with our credence in tails split evenly between Monday and Tuesday. Lewis' account diverges from my account in what happens when you learn that it's Monday. Lewis holds that you should be a centered conditionalizer and come to have a $\frac{2}{3}$ credence in heads and a $\frac{1}{3}$ credence in tails.

Consider a subject with more than one doxastic world, who undergoes a belief change which just increases or decreases the number of alternatives at a world (to a minimum of 1). As we'll see, we can capture the flavor of these three accounts by looking at how such a belief change affects the subject's credence in that world. On my account the subject's credence remains unchanged. On Lewis' account, if the number of alternatives at that world increases then the subject's credence will remain unchanged. But if the number of alternatives at that world decreases, then the subject's credence will decrease as well. On Elga's account the subject's credence will change in both cases. If the number of alternatives at that world increases or decreases, then the subject's credence in that world will likewise increase or decrease.

In the next two sections I'll look in more detail at how Elga and Lewis treat the sleeping beauty case. Before we do that, a caveat is in order. Neither Elga nor Lewis offer an explicit account of the dynamics of *de se* beliefs they endorse. So in presenting Elga's and Lewis' arguments I've had to add implicit premises that their arguments require. That said, I take the accounts I offer on their behalf to be fair.

5 Elga's Response to Sleeping Beauty

In Elga's (2000) account of the sleeping beauty case, he proposes that after waking up our credence in heads/tails should be $\frac{1}{3}/\frac{2}{3}$, the latter split evenly between Monday and Tuesday. If we then learn it's Monday, he proposes that our credence in heads/tails should become $\frac{1}{2}/\frac{1}{2}$. Elga's proposal follows from four principles:

1. Centered Conditionalization
2. The Principal Principle
3. Elga's Indifference Principle
4. A Continuity Principle

We've already looked at centered conditionalization, and the Principal Principle is familiar. The third principle, Elga's (2004) Indifference Principle, states that

your credences in doxastic alternatives at the same world should be equal.[17] The fourth principle is a Continuity Principle. As we've seen, the content of a Continuity Principle depends on when we take pairs of alternatives to be continuous. For Elga's proposal, any Continuity Principle that includes the following sufficient condition for continuity will do: a new and old alternative are continuous if (i) both are centered at the same world and individual, (ii) the new alternative is not centered at an earlier time than the old alternative, and (iii) there's no other new alternative satisfying (i) and (ii) that's centered at an earlier time than this new alternative.

Given these four principles, Elga's proposal follows. Let $cr(\cdot)$ be your credence function and $hp(\cdot)$ your hypothetical priors. Let H/T be the propositions that the coin came up heads/tails, and SUN/MON/TUE be the centered propositions that it's Sunday/Monday/Tuesday.

By the Principal Principle, your credences in your heads and tails alternatives on Sunday will be $cr(H\wedge SUN) = cr(T\wedge SUN) = \frac{1}{2}$. Given centered conditionalization, this entails that $hp(H\wedge SUN) = hp(T\wedge SUN)$. When you wake up on Monday, your Sunday alternatives are replaced by Monday alternatives at the heads worlds, and by Monday and Tuesday alternatives at the tails worlds. Both the Monday and the Tuesday alternatives are centered at the same worlds and individuals as the Sunday alternatives, and at later times. But the Monday alternatives are centered at an earlier time than the Tuesday alternatives. So according to the Continuity Principle given above, it's the Monday (not Tuesday) alternatives that are continuous with the Sunday alternatives. We saw in section 3 that given centered conditionalization, the Continuity Principle requires that the ratios of priors between the new and old continuous alternatives be the same. Since $hp(H\wedge SUN) = hp(T\wedge SUN)$, it follows that $hp(H\wedge MON) = hp(T\wedge MON)$. Elga's Indifference Principle requires that your credences in the two alternatives at the tails worlds be equal, and given centered conditionalization this entails that $hp(T\wedge MON) = hp(T\wedge TUE)$. Putting these equalities together, we get $hp(H\wedge MON) = hp(T\wedge MON) = hp(T\wedge TUE)$. When you wake up your doxastic possibilities are $H\wedge MON$, $T\wedge MON$ and $T\wedge TUE$, so on centered conditionalization your credences after waking on Monday are $cr(H\wedge MON) = cr(T\wedge MON) = cr(T\wedge TUE) = \frac{1}{3}$.

Now, say you're woken up at 9 am. What if at 9:01 am you learn that it's Monday? After learning it's Monday you will have one alternative at each world, $H\wedge MON(9{:}01)$ at the heads worlds and $T\wedge MON(9{:}01)$ at the tails worlds. According to the Continuity Principle, $H\wedge MON(9{:}01)$ and $T\wedge MON(9{:}01)$ are continuous with $H\wedge MON(9{:}00)$ and $T\wedge MON(9{:}00)$, respectively. And since $hp(H\wedge MON(9{:}00)) = hp(T\wedge MON(9{:}00))$, it follows that $hp(H\wedge MON(9{:}01)) = hp(T\wedge MON(9{:}01))$. So on centered conditionalization your credence after learning it's Monday is evenly split between heads and tails: $cr(H\wedge MON(9{:}01)) = cr(T\wedge MON(9{:}01)) = \frac{1}{2}$.

Note that the Principal Principle only plays a superficial role in the argument for Elga's proposal. The Principal Principle sets our credences in heads and tails

[17] Elga (2004) proposes that subjectively indistinguishable alternatives at the same world should be assigned the same credence. Assuming that one's current evidence includes one's current subjective state, it follows that all of one's alternatives are subjectively indistinguishable, and Elga's Indifference Principle becomes the claim that alternatives at the same world should have the same credences.

on Sunday to $\frac{1}{2}/\frac{1}{2}$. But the argument goes through equally well given any reason for $\frac{1}{2}/\frac{1}{2}$ credences in heads and tails on Sunday. Likewise, the argument goes through just as well if heads and tails are replaced by two different hypotheses we have other reasons for having $\frac{1}{2}/\frac{1}{2}$ credences in.

In the sleeping beauty case it's uncontentious that the Principal Principle applies on Sunday, and thus that you should have $\frac{1}{2}/\frac{1}{2}$ credences in heads and tails. Some of the sleeping beauty literature has focused on whether the Principal Principle should also apply after you wake up on Monday.[18] The question is whether you get admissible evidence when you wake up on Monday. If so, the thought goes, then the Principal Principle should still apply, and your credences in heads and tails should remain $\frac{1}{2}/\frac{1}{2}$.

It follows from Elga's argument that upon waking our credences in heads and tails should be $\frac{1}{3}/\frac{2}{3}$. So if Elga's argument is sound, you do get inadmissible evidence when you wake up on Monday. But I think debating admissibility and the Principal Principle is the wrong way to approach the problem. First, there is no agreement as to what counts as admissible evidence.[19] This makes it hard to make progress in a debate over whether someone's evidence is admissible. Second, focusing on the issue of whether the Principal Principle applies on Monday gets us relatively little. As we just saw, the argument goes through just as well if heads and tails are replaced by two different hypotheses we have other reasons for having $\frac{1}{2}/\frac{1}{2}$ credences in. Concluding one thing or another about the Principal Principle doesn't tell us what to say in these other cases. Finally, suppose we decide that we don't receive inadmissible evidence upon waking, and therefore that Elga's argument is incorrect. We still need to decide what part of Elga's argument to reject, since the argument entails the $\frac{1}{3}/\frac{2}{3}$ result without making any assumptions about the admissibility of your evidence on Monday. The argument only requires that the Principal Principle hold on Sunday, before you go to sleep. Given this, I think it's better to assess the merits of Elga's argument and then see what implications this has regarding admissibility than to use admissibility to assess the merits of Elga's argument.

If one accepts Elga's argument, belief changes that increase the number of doxastic alternatives at a world will generally increase one's credence in that world relative to worlds without such an increase. Likewise, one's credence in a proposition which multiplies doxastic alternatives will generally increase relative to propositions that don't multiply alternatives. One can see why this should be so for the proponent of Elga's response: to endorse Elga's response is to think that one's credence in tails should increase relative to one's credence in heads when the number of alternatives given tails increases (and the number of alternatives given heads does not).

However, accepting Elga's argument leads to counterintuitive consequences. Consider the following case:

> *The Many Brains Argument:* Consider the hypothesis that you're a brain in a vat. I take it that this is epistemically possible and (perhaps) nomologically possible. Your current credence in this possibility,

[18] See Lewis (2001) and Dorr (2002).

[19] Though see Hall (2004) and Meacham (2005) for proposals regarding admissibility.

> however, is presumably very low. Now consider the proposition that you're in a world where brains in vats are constantly being constructed in states subjectively indistinguishable from your own. Let your credence in this proposition be $0 < p < 1$, and your credence that there will be no multiplication of doxastic alternatives be $1 - p$. If you accept Elga's argument then your credence in this hypothesis should be constantly increasing and will converge to 1. Thus, if you hold such a position you should come to believe (if not yet, then in a little while) that these brains in vats are being created. (A proof of this result is provided in the appendix.)

It follows from Elga's Indifference Principle that your credences should be spread evenly among the doxastic alternatives at a world. So as you become certain that these brains in vats are being created, you should become certain that you're a brain in a vat.

The many brains argument assumed that brain in a vat duplication is the only proposition in which you have a non-zero credence that multiplies doxastic alternatives. Now suppose that you also have a small credence in the proposition that you're in a world where duplicates of you are constantly being created on distant but qualitatively identical planets. Then you'll come to believe (if not yet, then in a little while) that these brains in the vats are being created *or* that these duplicates of you are being created. Likewise, you'll come to believe that you are a brain in a vat *or* a duplicate on a distant planet. By a similar process, you can generalize the result of the many brains argument to any number of propositions that multiply alternatives.

In general, if you accept Elga's argument then you will come to believe that you're in a world where you have many doxastic alternatives. These are strange worlds. So if we accept Elga's argument, we'll come to believe (if not yet, then in a little while) that we live in a strange world. This is an unwelcome consequence.

6 Lewis' Response to Sleeping Beauty

In his criticism of Elga's account of sleeping beauty, Lewis (2001) claims that you do not receive inadmissible evidence when you wake up on Monday. Thus the Principal Principle should still apply on Monday, and your credence in heads/tails should remain $\frac{1}{2}/\frac{1}{2}$. I've said above why I think this is the wrong way to approach the problem. And as we saw, even if Lewis is right there remains the task of deciding what's wrong with Elga's argument, since the argument only requires that the Principal Principle apply on Sunday. So how would Lewis address Elga's argument? To reject the argument, Lewis needs to reject one of the four premises the argument employs. With Elga, Lewis accepts that the Principal Principle entails that our credences in heads and tails on Sunday should be $\frac{1}{2}/\frac{1}{2}$. Furthermore, Lewis endorses Elga's Indifference Principle and (centered) conditionalization. So Lewis must reject Elga's Continuity Principle.

In Lewis' (2001) account of the sleeping beauty case, he proposes that after waking up our credence in heads/tails should be $\frac{1}{2}/\frac{1}{2}$, the latter split evenly between Monday and Tuesday. If we then learn it's Monday, he proposes that our

credence in heads/tails should become $\frac{2}{3}/\frac{1}{3}$. Lewis' proposal follows from five principles:

1. Centered Conditionalization
2. The Principal Principle
3. Elga's Indifference Principle
4. A Continuity Principle
5. The No-Increase Principle

The first three premises are familiar. The fourth premise is another Continuity Principle. Although Lewis must reject Elga's Continuity Principle, we can use it to characterize a Continuity Principle that will suit Lewis' purposes. Elga's Continuity Principle requires that any old and new alternative that satisfy the following conditions be continuous: (i) both are centered at the same world and individual, (ii) the new alternative is not centered at an earlier time than the old alternative, and (iii) there's no other new alternative satisfying (i) and (ii) that's centered at an earlier time. Lewis' Continuity Principle requires that any pair of alternatives that satisfies these conditions be continuous *iff* the number of alternatives at that world has not increased.

Lewis' Continuity Principle needs to deny that pairs of alternatives that satisfy these conditions are continuous when the number of alternatives at that world increases. This leaves us with the question of what constraints, if any, should be imposed on your credences at worlds where the number of alternatives increases. Lewis' position seems to be that in cases where you don't suffer from memory loss and don't get evidence about the world—where you don't gain or lose doxastic worlds—increases in the number of alternatives at a world should leave your credence in that world unchanged. I'll call this the No-Increase Principle.

Given these five principles, Lewis' proposal follows. As before, the Principal Principle and centered conditionalization entail that hp(H∧SUN) = hp(T∧SUN). When you wake up on Monday your Sunday alternatives are replaced by Monday alternatives at the heads worlds and by Monday and Tuesday alternatives at the tails worlds. By the No-Increase Principle the increase in alternatives at your tails worlds leaves your credence in tails unchanged, so your credence in tails after waking up on Monday is the same as your credence in tails on Sunday, $\frac{1}{2}$. So your credence in heads after waking up on Monday must be $\frac{1}{2}$ as well. Given centered conditionalization, this entails that hp(H∧MON) = hp(T∧(MON∨TUE)) = hp(T∧MON) + hp(T∧TUE). Elga's Indifference Principle and centered conditionalization entail that hp(T∧MON) = hp(T∧TUE). Taken together, these equalities entail hp(H∧MON) = hp(T∧MON) + hp(T∧TUE) = 2·hp(T∧MON) = 2·hp(T∧TUE). When you wake up your doxastic possibilities are H∧MON, T∧MON and T∧TUE, so on centered conditionalization your credences after waking up on Monday are cr(H∧MON) = $\frac{1}{2}$ and cr(T∧MON) = cr(T∧TUE) = $\frac{1}{4}$.

Now what if you're woken up at 9 am and told at 9:01 am that it's Monday? After learning it's Monday you will have one alternative at each world, and by the Continuity Principle these alternatives will be continuous with your Monday 9 am alternatives. We know from above that hp(H∧MON(9:00)) = 2·hp(T∧MON(9:00)), so it follows that hp(H∧MON(9:01)) = 2·hp(T∧MON(9:01)).

So on centered conditionalization your credences after learning it's Monday are $\text{cr}(\text{H}\wedge\text{MON}(9{:}01)) = \frac{2}{3}$ and $\text{cr}(\text{T}\wedge\text{MON}(9{:}01)) = \frac{1}{3}$.

Elga's account ran into problems because it entailed that belief changes that multiply alternatives at a world generally increase one's credence in that world. Lewis avoids this result by adopting a different Continuity Principle and the No-Increase Principle. But while on Lewis' account belief changes that multiply alternatives at a world don't increase one's credence in that world, belief changes that decrease the number of alternatives at a world generally do decrease one's credence in that world. And this leads to counterintuitive consequences for his account as well. Consider the following case:

> *The Sadistic Scientists Argument:* Consider the hypothesis that you're in a world where every second some scientists will create n brains in vats in situations subjectively identical to your own. A half second after the brains are created, the scientists will destroy them. Let your credence in this proposition be $0 < p < 1$, and your credence that there will be no creation or destruction of doxastic alternatives be $1 - p$. When the brains are created your credence that you are in such a world will remain the same (No-Increase Principle), and this credence will be evenly split between your $n + 1$ alternatives (Indifference Principle). As a half second passes and these brains are destroyed, your credence that you are in such a world will decrease by the appropriate amount (Continuity Principle and centered conditionalization). So as each second passes, your credence that you are in such a world will decrease and converge to 0. Thus, if you hold Lewis' position you should come to believe (if not yet, then in a little while) that these brains in vats are not being created. (A proof of this result is provided in the appendix.)

The sadistic scientists argument assumed that brain in vat destruction is the only proposition you have a non-zero credence in that diminishes alternatives. Now suppose that you also had a small credence in the proposition that duplicates of you on distant but qualitatively identical worlds were being created and destroyed. Then you'd come to believe (if not yet, then in a little while) that neither of these propositions was true. The result generalizes to any number of propositions that diminish alternatives. In general, if you accept Lewis' argument then you'll come to believe that you're not in a world where continual doxastic elimination is taking place.

I take this result to be counterintuitive. If the result as stated does not move you, imagine a case in which you are living in a world where brain-in-a-vat creation technology is cheap and easily accessible. An enemy of yours who would enjoy destroying brains in vats in your subjective state tells you that at midnight she'll spend an hour creating n such brains, and at 1 am she'll spend an hour destroying them. This enemy has the resources to carry out this threat, and reliably carries out the threats she makes. If n is big enough, and you uphold the Lewis' account, then though you're now almost certain that she will carry out her threat, when you wake up tomorrow morning you'll be almost certain that she didn't. Indeed, if n is big enough, you could even go with her and watch as she creates the brains and destroys them; if you watch for long enough you won't believe your eyes!

7 The Varied Brains Argument

In the last two sections I've argued that Elga's and Lewis' positions lead to counterintuitive consequences. Now let's turn a critical eye toward my account.

Consider a case like sleeping beauty, but with the following twist. If the coin toss comes up tails, the scientists will put you in a black room on Monday and a white room on Tuesday. If the coin toss comes up heads, they'll flip another coin to determine whether to put you in a black or white room on Monday.

What should your credences be in this case on the three accounts we've looked at? On all three accounts your credences in heads and tails on Sunday will be the same as in the sleeping beauty case. Likewise, on all three accounts your credences in heads and tails after waking up on Monday before you open your eyes will be the same as in the sleeping beauty case. What about your credences in heads and tails after you open your eyes and see a black room? On Elga's and Lewis' accounts your credences will be the same: half of the heads worlds are eliminated and half of the tails alternatives are eliminated, and after renormalizing you get the same credences in heads and tails as before. Not so for compartmentalized conditionalization. Half of the heads worlds are eliminated, but none of the tails worlds are (half of the tails *alternatives* are eliminated, but this doesn't eliminate any of the tails worlds) and after renormalizing your credence in tails will go up. So when you open your eyes and see a black room, your credence in heads/tails will become $\frac{1}{3}/\frac{2}{3}$.

This raises a natural worry for my account. I offered the many brains argument as a criticism of Elga's $\frac{1}{3}/\frac{2}{3}$ response to the sleeping beauty case. In the black and white room version of sleeping beauty compartmentalized conditionalization also ends up assigning $\frac{1}{3}/\frac{2}{3}$ credences to heads and tails. Is there an argument analogous to the many brains argument against compartmentalized conditionalization?

Yes and no. Let's look at how such an argument might go. The many brains argument itself won't work because on compartmentalized conditionalization multiplying alternatives at a world doesn't increase the likelihood of that world. As long as our doxastic worlds remains the same, our credences in worlds will remain the same. To get an argument analogous to the black and white room case, we need an argument where the normal worlds are eliminated but the alternative multiplying worlds are not. So consider the following case:

> *The Varied Brains Argument:* Assume that your doxastic worlds are such that they can be divided into two kinds of worlds, normal worlds and strange worlds. Throughout your doxastic worlds there are n subjectively distinguishable experiences that you might experience in the next second. Assume that you have some normal doxastic world compatible with each experience, and you have no subjective duplicates at your normal doxastic worlds. Assume that at each of your strange doxastic worlds there are scientists that will create n brains in vats a second from now, each brain compatible with one of your possible experiences. Now, at the end of a second you'll have some experience, say that of eating chocolate ice cream. This will eliminate the many normal worlds in which you don't have the experience of eating chocolate ice cream. On the other hand, at all of your strange worlds there's a brain in a vat

> which has the experience of eating chocolate ice cream, so no strange worlds will be eliminated. On compartmentalized conditionalization, your credence in your strange doxastic worlds should increase relative to your credence in your normal doxastic worlds.

We can extend this case by replacing 'second' with longer units of time, and as the unit of time grows larger, the number n of distinguishable experiences you might experience grows larger as well. By making the unit of time arbitrarily large, we can get a case in which, on compartmentalized conditionalization, your credence in your strange doxastic worlds grows arbitrarily large.

How bad is this?

One might question whether this result is counterintuitive. This is an interesting, if murky, question. But it is worth looking at how things stand if we decide that the result is counterintuitive.

In the varied brains case, your credence in your strange worlds increases relative to your credence in your normal worlds because of the artificial way in which these doxastic worlds have been selected: all the strange worlds under consideration are ones that will end up matching whatever you experience, whereas many of your normal worlds won't match what you experience. If we restricted the normal worlds to those compatible with eating chocolate ice cream, your credence in your strange worlds would not increase relative to your credence in your normal worlds. Likewise, if we placed no restrictions on which strange worlds were allowed, then the experience of eating chocolate ice cream would eliminate lots of strange worlds as well as lots of normal worlds. Whether your credence in strange worlds increases relative to your credence in normal worlds depends on which strange and normal worlds are your doxastic worlds—which worlds our priors and evidence lead us to believe could be ours. And it's reasonable to think that if you have doxastic worlds like ours, your credence in strange worlds will not gain on your credence in normal worlds.

Skeptical results can be roughly divided into two kinds. First, there are results which entail that people like us in situations like ours should be lead to skepticism. Second, there are results which entail skeptical consequences for people in outlandish situations, but which have little bearing on people like us. I take it that the first kind of result is worse than the second. Our general sentiment is that our intuitions in outlandish situations are less reliable—and thus easier to discard—than our intuitions in situations we're familiar with. Likewise, it's easier to bite the bullet with counterintuitive cases that have little impact on our everyday lives.

The varied brains argument is a result of the second kind; it entails that people with certain idiosyncratic doxastic set-ups will come to believe something counterintuitive. The many brains argument, on the other hand, is a result of the first kind; it entails that people like us should come to believe that we live in a strange world. So the skeptical arguments considered weigh more heavily against Elga's account than they do against the account I favor.

What about the sadistic scientists argument? This too is a result of the second kind. While people like us will become more and more sure we're not in a 'diminishing' world, this will have little effect on overall belief distribution since our credences in such possibilities are so small. Only people whose initial credence in these strange worlds are high will be lead to counterintuitive results. So the

skeptical arguments, considered in isolation, don't leave us with a reason to favor the account I advocate over Lewis' account. It is other considerations, such as the *prima facie* plausibility of the view, the implications with regards to reflection and continuity, etc., that will decide between the two views.

8 A Third Account

We can sum up the intuitive difference between the three accounts with the following case:

> *The Up-and-Down Case:* Suppose you learn that you'll be part of the following experiment. Some scientists will flip a fair coin tonight. If it comes up tails, then every day at noon the scientists will create n brains in vats in states subjectively identical to yours, and at midnight will destroy $\frac{n}{2}$ of them. If it comes up heads, no brains will be created or destroyed.

If you endorse Elga's account then your credence that the coin came up tails will converge to 1, regardless of your evidence (knowledge of objective chances, etc.) to the contrary. If you endorse Lewis' account then your credence that the coin came up heads will converge to 1, regardless of your evidence (knowledge of objective chances, etc.) to the contrary. If you endorse my account, then your credences in heads and tails will remain $\frac{1}{2}/\frac{1}{2}$.

In this paper I've offered two reasons to adopt the third option. First, the dynamics my account employs has a substantial advantage over the dynamics of Elga's and Lewis' with regards to accommodating the continuity of our beliefs. Second, I've shown that while all three accounts arguably suffer from counterintuitive consequences, the consequences faced by my account are better than those faced by Elga's account, and no worse than those faced by Lewis' account.

The preceeding discussion allows us to see the first reason for favoring my account in a different light. As we saw in section 3, the fact that compartmentalized conditionalization does better at capturing our intuitions regarding continuity can be cashed out in terms of priors: the Continuity Principle requires a much weaker constraint on priors if we adopt compartmentalized conditionalization than if we adopt centered conditionalization. We can see how weak this constraint is by noting that, given compartmentalized conditionalization, Elga's Indifference Principle imposes a strictly stronger constraint on priors than the Continuity Principle. That is, if you satisfy Elga's Indifference Principle then you'll automatically satisfy the Continuity Principle as well.[20] This gives those who find Elga's Indifference Principle independently plausible a further reason to favor compartmentalized conditionalization: given compartmentalized conditionalization, a person who adopts

[20]Recall that given compartmentalized conditionalization, the Continuity Principle requires that the ratio of priors between new alternatives at each world be the same as the ratio of priors between any old alternatives at that world that they're continuous with. If you adopt Elga's Indifference Principle, then your credences in alternatives at a world will be the same, and thus so will your priors. If your priors in alternatives at a world are always the same, the ratio of priors between alternatives at a world will always be 1:1, and the Continuity Principle will be satisfied.

Elga's Indifference Principle needn't adopt any further principles in order to get completely coordinated credences.

Compartmentalized conditionalization provides another advantage as well. We saw in section 3.1 that it's hard to give a precise characterization of when alternatives are continuous. If we adopt centered conditionalization, then providing such a characterization is urgent if we're to know how to coordinate our credences. If we adopt compartmentalized conditionalization, this is far less urgent. At most of our doxastic worlds—worlds with a single alternative—it doesn't matter, since our alternatives at these worlds will automatically have coordinated credences, regardless of whether they're continuous or not. If we adopt compartmentalized conditionalization and Elga's Indifference Principle, we don't need to provide such a characterization at all: our alternatives at *all* worlds will automatically have coordinated credences.[21]

These aren't the only considerations relevant to the assessment of these three accounts. There are further questions about betting arguments, reflection principles, and the like. But if what I've said is right, these two considerations provide compelling reasons in favor of my account.[22]

[21]To see this note that if we adopt compartmentalized conditionalization and Elga's Indifference Principle all of our alternatives will satisfy the priors constraint that the Continuity Principle requires of only continuous alternatives.

[22]For valuable comments and discussion, I'd like to thank Frank Arntzenius, Maya Eddon, Adam Elga, Hilary Greaves, John Hawthorne, David Manley, Tim Maudlin, Adam Sennet, Ted Sider, Jonathon Weisberg and an anonymous referee. In particular, I owe much to Tim Maudlin, whose comments on Elga's account and the many-worlds interpretation of quantum mechanics inspired my interest in these issues, and David Manley, for raising the black and white room case. Finally, I owe a special thanks to Frank Arntzenius, Maya Eddon, and John Hawthorne for comments on several of drafts of this paper.

A The Many Brains Argument

For simplicity, assume that there are only two worlds under consideration, one normal world and one brain-duplicating world; it's easy to see how the result generalizes to multiple worlds. Let S be the stable world, and D be the duplicating world.

Consider the alternatives focused on the original (non-brain) individual at the S and D worlds. As time changes you will replace these alternatives with new alternatives at those worlds, centered on the same individual and a later time. (At the D world, of course, you will also be replacing old brain-centered alternatives with their temporal successors, as well as adding entirely new brain alternatives.) The new non-brain alternatives and the old non-brain alternatives they replaced satisfy the three conditions of Elga's Continuity Principle. We saw in section 3 that given centered conditionalization, the Continuity Principle requires that the ratios of priors between new and old continuous alternatives be the same. So the ratio of your priors in the non-brain alternatives at the D and S worlds at a time will be constant. That is, if we let pr_{S_t} and pr_{D_t} be your priors in the non-brain alternatives at the D and S worlds at t, the Continuity Principle entails that $\forall t \left(\frac{\mathrm{pr}_{D_t}}{\mathrm{pr}_{S_t}} = k\right)$, for some constant k.

Elga's Indifference Principle entails that one's credences in alternatives at a world be the same, and thus (given centered conditionalization) that one's priors in alternatives at a world be the same. So one's prior in the brain alternatives centered on world D and time t will be the same as your prior in the non-brain alternative centered on world D and time t, pr_{D_t}.

Now, let N_{W_t} be the number of alternatives you have at time t that are centered on a world W, and let $\mathrm{cr}_t(W)$ be your credence at t in W. Assume the brains are created one at a time, and choose temporal units and a temporal origin such that (a) $\mathrm{N}_{D_0} = \mathrm{N}_{S_0} = 1$, and (b) $\mathrm{N}_{D_t} = t{+}1$. Since you only ever have one alternative centered on S, $\forall t\ (\mathrm{N}_{S_t} = 1)$.

Centered conditionalization and the above then entail that:

$$\begin{aligned}
\mathrm{cr}_t(D) &= \frac{N_{D_t} \cdot \mathrm{pr}_{D_t}}{N_{D_t} \cdot \mathrm{pr}_{D_t} + N_{S_t} \cdot \mathrm{pr}_{S_t}} \\
&= \frac{N_{D_t} \cdot \mathrm{pr}_{D_t}}{N_{D_t} \cdot \mathrm{pr}_{D_t} + N_{S_t} \cdot \frac{\mathrm{pr}_{D_t}}{k}} \\
&= \frac{N_{D_t}}{N_{D_t} + \frac{N_{S_t}}{k}} \\
&= \frac{t+1}{t+1+\frac{1}{k}}.
\end{aligned}$$

Thus:

$$\lim_{t\to\infty} (\mathrm{cr}_t(D)) = \lim_{t\to\infty} \left(\frac{t+1}{t+1+\frac{1}{k}}\right) = 1.$$

B The Sadistic Scientists Argument

Again, for simplicity assume that there are only two worlds under consideration, one normal world and one brain-duplicating-and-destroying world. Let S be the stable world, and D be the duplicating-and-destroying world.

As before, let N_{W_t} be the number of alternatives you have at time t that are centered on a world W, and let $\mathrm{cr}_t(W)$ be your credence at t in W. Choose temporal units and a temporal origin such that if $t < 0$ or $t > n$, then $N_{D_t} = 1$, and if $0 \leq t \leq n$, then $N_{D_t} = (n+1) - t$. (So n is the number of brains that will be created in D at time $t = 0$, and one of these brains will be destroyed every unit of time thereafter.)

As before, let pr_{S_t} and pr_{D_t} be your priors in the non-brain alternatives at the D and S worlds at t. Now consider the alternatives focused on the original (non-brain) individual at the S and D worlds. As time changes you will replace these alternatives with new alternatives at those worlds, centered on the same individual and a later time. (At the D world, of course, you will also be replacing old brain-centered alternatives with their temporal continuants, as well as adding entirely new brain alternatives.) The new non-brain alternatives and the old non-brain alternatives they replaced satisfy the conditions of Lewis' Continuity Principle until time $t = 0$, when the brains are created. So the continuity principle entails that for $t < 0$, $\left(\frac{\mathrm{pr}_{D_t}}{\mathrm{pr}_{S_t}} = k\right)$, for some constant k. The conditions also hold after the brains are created, so the continuity principle entails that for $t \geq 0$, $\left(\frac{\mathrm{pr}_{D_t}}{\mathrm{pr}_{S_t}} = l\right)$, for some constant l.

Elga's Indifference Principle entails that one's credences in alternatives at a world be the same, and thus (given centered conditionalization) that one's priors in alternatives at a world be the same. So one's prior in the brain alternatives at D at t will be the same as your prior in the non-brain alternative at D, pr_{D_t}. The No-Increase Principle entails that your credence in D shouldn't change when the new brains are created at $t = 0$. This, centered conditionalization and the above entail that $l = k/(n+1)$.

Centered conditionalization and the above then entail that:

$$\begin{aligned}
\mathrm{cr}_{t=n}(D) &= \frac{N_{D_n} \cdot \mathrm{pr}_{D_n}}{N_{D_n} \cdot \mathrm{pr}_{D_n} + N_{S_n} \cdot \mathrm{pr}_{S_n}} \\
&= \frac{N_{D_n} \cdot \mathrm{pr}_{D_n}}{N_{D_n} \cdot \mathrm{pr}_{D_n} + N_{S_n} \cdot \mathrm{pr}_{D_n} \cdot (n+1)/k} \\
&= \frac{\mathrm{pr}_{D_n}}{\mathrm{pr}_{D_n} + \mathrm{pr}_{D_n} \cdot (n+1)/k} \\
&= \frac{1}{1 + (n+1)/k}.
\end{aligned}$$

Thus:

$$\lim_{n \to \infty} \left(\mathrm{cr}_{t=n}(D)\right) = \lim_{n \to \infty} \left(\frac{1}{1 + \frac{n+1}{k}}\right) = 0.$$

References

Arntzenius, F. (2002) "Reflections on Sleeping Beauty", *Analysis*, 62: 53-61

Arntzenius, F. (2003) "Self-locating Beliefs, Reflection, Conditionalization and Dutch Books", *Journal of Philosophy*, 100: 356-370

Bartha, P. and Hitchcock, C. (1999) "No one knows the date or the hour: an unorthodox application of Rev. Bayes' Theorem", *Philosophy of Science (Proceedings)*, S339-353

Dorr, C. (2002) "Sleeping Beauty: in defense of Elga", *Analysis*, 62: 292-296

Earman, J. (1992) *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*, MIT Press

Elga, A. (2000) "Self-locating belief and the Sleeping Beauty problem", *Analysis*, 60: 143-147

Elga, A. (2004) "Defeating Dr. Evil with self-locating belief",*Philosophy and Phenomenological Research*, 69: 383-396

Hall, N. (1994) "Correcting the Guide to Objective Chance", *Mind*, 103: 505-517

Hall, N. (2004) "Two Mistakes About Credence and Chance", *Australasian Journal of Philosophy*, 82: 93-111

Halpern, J. (2004) "Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems", *Proceedings of the Twentieth Conference on Uncertainty in AI*, 226-234

Hitchcock, C. (2004) "Beauty and the Bets", *Synthese*, 139: 405-420

Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach, 2nd Ed.*, Open Court Publishing Company

Lewis, D. (1979) "Attitudes *De Dicto* and *De Se*" in *The Philosophical Review*, 88: 513-543

Lewis, D. (1980) "A Subjectivist's Guide to Objective Chance" in *Studies in Inductive Logic and Probability, Vol. 2*, edited by Richard C. Jeffrey, University of California Press

Lewis, D. (2001) "Sleeping Beauty: reply to Elga", *Analysis*, 61: 171-176

Maher, P. (2005) "The Concept of Inductive Probability", *Erkenntnis*, forthcoming

Meacham, C. (2005) "Three Proposals Regarding a Theory of Chance", *Philosophical Perspectives*, 19: 281-307

Strevens, M. (2004) "Bayesian Confirmation Theory: Inductive Logic, or Mere Inductive Framework?", *Synthese*, 141: 365-379