Sleeping Beauty and the Dynamics of *De Se* Beliefs

Christopher J. G. Meacham

1 Introduction

Take beliefs to be narrowly psychological. Then there are two types of beliefs.[1] First, there are beliefs about what the world is like, or *de dicto* beliefs. Taking a proposition to be a set of possible worlds, the objects of *de dicto* beliefs are propositions. To believe a proposition is to believe that your world is one of the worlds that form that proposition. 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 one of these worlds.

But not all beliefs are beliefs in propositions. Take a world where two wise sages live, Zorn and Xingu. Both sages know which world they're in, and thus which propositions are true. The two sages live on different planets, but the planets are qualitatively identical. Furthermore, the sages themselves are qualitatively identical.

Now, both sages can't have true beliefs about which sage they are. The two sages are qualitatively identical, so if Zorn believes he is Zorn, so does Xingu. If both sages believe that they're Zorn, then Xingu has a false belief, even though all his propositional

[1] I borrow liberally here from David Lewis (1979). In addition to assuming beliefs are in the head, I'll follow Lewis in ignoring difficulties that arise from mathematical or logical truths, and in assuming that the subjects of belief attitudes only exist at one world. I employ Lewis' framework for its elegance, but I think most of my substantive points don't depend on it.

In particular, note that nothing depends on the outcome of the internalist/externalist debate. It might be that the best candidate for the meaning of 'belief' is one where beliefs aren't in the head, as the externalists claim. In any case, there is also another, if less eligible, candidate for the meaning of 'belief' where beliefs are in the head. Call the first candidate belief$_1$, the second belief$_2$, and take me to be talking about beliefs$_2$.

beliefs are true. So not all beliefs are propositional. Over and above beliefs about what the world is like, there are beliefs about where one is in the world.

Beliefs broadly construed are *de se* beliefs. A *centered world* is a possible world paired with a designated individual and a time. A *centered proposition* is a set of centered worlds. The objects of *de se* beliefs are centered propositions. To believe a centered proposition is to believe that your centered world—who and when you are, and in what world—is one of the centered worlds that form that centered proposition.

We can turn any proposition or set of worlds into an equivalent centered proposition or set of centered worlds, by replacing each world with all the centered worlds at that world. Thus all *de dicto* beliefs are reducible to *de se* beliefs. Of course, not all *de se* beliefs are reducible to *de dicto* beliefs. *De se* beliefs that aren't reducible to *de dicto* beliefs are *self-locating* or *irreducibly de se* beliefs.

In his influential article "Attitudes *De Dicto* and *De Se*", David Lewis asks what happens to Bayesian decision theory once we consider self-locating beliefs as well as *de dicto* beliefs. Lewis' answer:

> "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]

I will argue that Lewis is mistaken. I think there is a deep divide between our beliefs about the world and our beliefs about our place in the world. I will argue that changes in one's purely self-locating beliefs should have no affect on one's *de dicto* beliefs. Moreover, I'll argue that this division should be a consequence of the dynamics we adopt for *de se* beliefs. The dynamics I advocate are essentially identical to those proposed by Halpern and Tuttle (1993) and Halpern (2004). (Some minor differences between our approaches are described in the following footnote.[3]) So this paper can be seen as providing further reasons for adopting their account.

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

[3]Unfortunately, the bulk of this work was done before I became aware that Halpern and Tuttle had already proposed essentially the same view. As a result, our presentations of this material differ in a number of ways. Likewise, the terminology used in this paper is not the same as that used in Halpern and Tuttle (1993) and Halpern (2004). Regarding the dynamics of *de se* beliefs, the account laid out here differs from theirs in the following minor respects: 1. Halpern and Tuttle use ordered pairs of a world and time in place of the Lewisian centered worlds used here (ordered triples of a world, time and individual). Thus formulated, their work does not apply to non-temporal cases of self-location, such as

My arguments for this conclusion will draw on the recent literature on the sleeping beauty problem. The sleeping beauty problem raises exactly the question of how changes in self-locating beliefs should affect our beliefs about the world. I'll show that two of the responses to the sleeping beauty problem that have been advocated in the literature lead to highly counterintuitive consequences. In light of this, I'll argue that we should adopt the account offered by Halpern and Tuttle.

The paper will proceed as follows. In the next section I'll present two competing dynamics for *de se* beliefs, the first employed by Elga and Lewis, the latter by Halpern and Tuttle. In the third and fourth sections I'll discuss some preliminary material needed for the discussion ahead. In the third section I discuss some of the continuity issues that arise in *de se* contexts, and in the fourth section I discuss purely self-locating beliefs. In the fifth section I'll present the sleeping beauty problem and sketch the three responses to it. In the sixth, seventh and eighth sections I'll look at Elga's and Lewis' responses in detail, and show how they both lead to highly counterintuitive consequences. In the ninth section I'll briefly look at some further considerations for and against these positions. I conclude in section ten.

2 The Big Picture

It's standard to assume that belief is not an all or nothing affair, but rather admits of degrees. A subject's beliefs are then represented by a credence 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 one of those possibilities obtains. The values it assigns are additive: the value it assigns to the union of several non-overlapping regions of the space is the sum of the values it assigns to each of these regions. The value it assigns to the entire space of possibilities is one, representing the subject's certainty that some possibility obtains.

In the case of *de dicto* beliefs, the space of possibilities is the space of possible worlds.

can happen in cases of duplication, fission, etc. 2. In the dynamics I present in section two, a subject's credences in doxastic alternatives are completely determined by her priors. The dynamics presented by Halpern and Tuttle are slightly less ambitious, and do not dictate a means by which one's credence in a doxastic world should be divided among the doxastic alternatives at that world.

The credence 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 each world in that proposition. The worlds in which the subject has non-zero credences are the worlds she thinks might be hers, or her *doxastic worlds*.

When we generalize to *de se* beliefs, the space of possibilities becomes the space of centered worlds. The credence 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 each centered world 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*.

Return to *de dicto* beliefs. On a broadly Bayesian picture, something like conditionalization will govern rational belief change.[4] Taking Earman's (1986) version of Bayesianism as a model, we can characterize *de dicto* conditionalization as follows.

A rational subject's credences are fixed by her hypothetical priors and her total evidence. A subject's credences are represented by a dynamic probability function, a function that changes with her evidence. A subject's hypothetical priors are represented by a static probability function, a function that encodes her disposition to respond to evidence. (Hypothetical priors are called 'priors' because they can be thought of as a rational subject's original credences in possibilities, prior to the receipt of any evidence, and 'hypothetical' because it is unlikely that one ever was in such a state.) A piece of evidence is represented by a proposition, and a subject's total evidence is represented by the conjunction of her evidential propositions. If a subject is rational, all belief changes will be the result of the addition of evidence.

A rational subject's credences can be determined from her priors and evidence directly, but it's convenient for our purposes to break this entailment into two steps. First, a rational subject's total evidence and hypothetical priors determine her doxastic worlds: her doxastic worlds are the worlds she has non-zero priors in that are compatible with her evidence. Second, a rational subject's hypothetical priors and doxastic worlds determine her credences. Her credences in non-doxastic worlds are, of course, zero. Her

[4]For simplicitly I'm ignoring Jeffery conditionalization and the like throughout this paper.

credences in doxastic worlds are obtained by normalizing her hypothetical priors in these worlds. That is, by assigning credences to each doxastic world such that they sum to one, and such that the ratios between her credences in these worlds are the same as the ratios between her hypothetical priors in these worlds.

We can see the effects of these constraints visually. Picture a subject's credence function as a three dimensional map, with each point on the plane representing a world, and the height at each point her credence in that world. There will be a boundary on this map outside of which everything is flat. This boundary outlines the set of doxastic worlds. Since a subject's belief in all possibilities sums to a constant, the volume inside the boundary is conserved. The relative height of the points inside the boundary is set by the subject's priors—numbers written at each point. The actual height of these points is then fixed by the set of doxastic worlds—which points are inside the boundary—which determines how thinly the volume inside the boundary is spread. Since priors are static, all belief changes are changes in the boundary. As the boundary shrinks, the points inside the boundary grow proportionally taller. (Since all rational belief changes are the result of added evidence, the boundary can only shrink, not expand.)

How should we generalize conditionalization to *de se* beliefs? One option is to replace every occurrence of 'world', 'proposition' and 'doxastic world' in the characterization of *de dicto* conditionalization just given with 'centered world', 'centered proposition' and 'doxastic alternative'. Seen visually, the picture will be just the same as before, except that each point on the map now represents a centered world instead of a world, and the boundary outlines the subject's doxastic alternatives instead of her doxastic worlds.

But Frank Arntzenius (2003) and Chris Hitchcock (2004) have shown that this version of *de se* conditionalization is untenable. Say you're looking at a clock you know to be accurate. The clock reads 6 pm, so your current credence that it's 6 pm is 1, and your credence that it's 6:01 pm is 0. A minute later the clock reads 6:01 pm, and your credence that it's 6 pm is 0, while your credence that it's 6:01 pm is 1. This violates the requirement that all belief changes be the result of the *addition* of evidence. The addition of evidence can only eliminate doxastic alternatives. But seeing the clock change did not just eliminate the doxastic alternatives where it's 6 pm, it also added doxastic alternatives at which it's 6:01 pm. To accommodate these kinds of cases we need to allow

a rational subject to both add and eliminate doxastic alternatives. (Arntzenius (2003) gives us good reason to revise *de dicto* conditionalization in the same way, allowing a rational subject to both add and eliminate doxastic worlds.[5] From now on I'll use '*de dicto* conditionalization' to refer to this appropriately modified version of conditionalization. In section nine I'll briefly need to speak of both the modified and unmodified versions; in that case I'll call them 'revised *de dicto* conditionalization' and 'unvrevised *de dicto* conditionalization', respectively.)

Take the version of *de se* conditionalization just considered, and relax the requirement that all rational belief changes be the result of the addition of evidence. Call this new version of *de se* conditionalization *centered conditionalization*. Seen visually, the picture is the same as it was before, except that now the boundary can both expand and contract.

Centered conditionalization is one way to generalize *de dicto* conditionalization to *de se* beliefs. You can't endorse both centered conditionalization and *de dicto* conditionalization, however, since their assignments conflict. 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 between alternatives, so that her credence in each alternative is $\frac{1}{4}$ and her credence in each world is $\frac{1}{2}$.[6] Now, what should her credences in A and B be if one of her alternatives at A is eliminated? According to *de dicto* conditionalization her credences in A and B should remain $\frac{1}{2}/\frac{1}{2}$. She has

[5]It's well known that there are cases where it appears that conditionalization is violated, such as cases of brainwashing or memory loss. These cases are usually circumvented by labeling belief changes that result from cognitive defects or memory loss 'irrational'. But Arntzenius (2003) presents us with a case where this move isn't available. In this case you flip a coin to see which of two routes you'll take to Shangri-la. If the coin comes up heads you'll travel by the mountains, if it comes up tails you'll travel by the sea. You further know that if you travel by the sea, then the guardians of Shangri-la will erase your memories of the trip when you arrive, and replace them with memories of having traveled by the mountains. Now say the coin comes up heads. While you're traveling by the mountains your credence in heads is 1. When you arrive at Shangri-la, however, it seems your credence in heads should become $\frac{1}{2}$, a violation of conditionalization. And this is so despite that fact that you've suffered from no memory loss or cognitive defects. It is the counterfactual possibility that you would have suffered from a memory loss had the coin come up tails that compels you to revise your credences, and this doesn't seem to provide grounds for the charge of irrationality.

[6]It follows from the additive nature of credences that a subject's credence in a world is equal to the sum of her credences in the centered worlds at that world (and, likewise, that a subject's prior in a world is equal to the sum of her priors in the centered worlds at that world).

the same doxastic worlds, so *de dicto* 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 her credences 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's another way to generalize conditionalization to *de se* beliefs which doesn't conflict with *de dicto* conditionalization. I'll call it *compartmentalized conditionalization*. Compartmentalized conditionalization is the same as centered conditionalization except that a different rule is used to determine a subject's credences given her hypothetical priors and doxastic alternatives. On centered conditionalization, the subject's priors in her doxastic alternatives are normalized. On compartmentalized conditionalization, the subject's hypothetical priors in her doxastic worlds are normalized, and then the subject's hypothetical priors in her doxastic alternatives are normalized at each doxastic world. That is, credences are assigned to each doxastic world such that they all sum to one, and such that the ratios between her credences in these worlds are the same as the ratios between her hypothetical priors in these worlds. Credences are then assigned to the doxastic alternatives at a world such that they sum to the credence allocated to that world, and such that the ratios between her credences in these alternatives are the same as the ratios between her hypothetical priors in these alternatives.

Again we can imagine this visually on a three dimensional map, where every point on the plane is a centered world and the boundary outlines the doxastic alternatives. As with *de dicto* conditionalization and centered conditionalization, on compartmentalized conditionalization all belief changes are changes in the boundary. However, the same boundary change will bring about different belief changes on compartmentalized conditionalization than it will on centered conditionalization. These differences are especially easy to visualize if we assume a subject has a finite number of alternatives, with equal priors in each. In this special case the volume inside the boundary will behave like a body of water. On centered conditionlization, imagine the boundary as a wall surrounding this body of water. As the boundary contracts, the water level rises, as the boundary expands, the water level falls. On compartmentalized conditionalization, imagine the

boundary as a wall surrounding a body of water, but this time with a number of inner walls dividing the body into cells. As the boundary shrinks only the water level in cells that are being contracted rise; the water level in the other cells will be unaffected. The exception is during the last step of contraction when a cell is eliminated. In this case the water in the cell is squeezed out over the cell walls and funneled into the surviving cells. Likewise, as the boundary grows, only the water level in cells that are expanding will fall; the water level in the other cells will be unaffected. The exception is when the boundary grows to the extent that a new cell is created, in which case water is funneled from all the other cells into it.[7]

I take David Lewis, Adam Elga and most of the sleeping beauty literature to endorse centered conditionalization. Joseph Halpern, Mark Tuttle and I endorse compartmentalized conditionalization.

3 Continuity

The dynamics of *de se* beliefs raises questions about belief continuity which don't arise in *de dicto* contexts. Consider again the case presented in the last section, where a subject is watching a clock they know to be accurate. When the clock changes from 6 pm to 6:01 pm, the subject discards all of her alternatives at which it's 6 pm and replaces them with alternatives at which it's 6:01 pm. Intuitively, her credences in these new alternatives should bear some relation to what her credences were 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 6 pm her priors in her two alternatives ($A_{6\text{pm}}$ and $B_{6\text{pm}}$) are equal, so her credences in $A_{6\text{pm}}$ and $B_{6\text{pm}}$ are $\frac{1}{2}/\frac{1}{2}$. When she sees the clock register 6:01 pm, what should her credences in $A_{6:01\text{pm}}$ and $B_{6:01\text{pm}}$ be? *Prima facie*, we have no reason to think they'll be $\frac{1}{2}/\frac{1}{2}$. Her priors in $A_{6\text{pm}}$ and $B_{6\text{pm}}$ were equal, but it's now 6:01 pm and these are no longer her

[7] To make this easy to visualize I've implicitly assumed that there's a single continuous boundary; i.e., assumed that the boundaries won't contract or expand in a way that forms islands.

alternatives. Her alternatives are now $A_{6:01\text{pm}}$ and $B_{6:01\text{pm}}$, and there's no reason that her priors in these alternatives should be equal.

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 change brings an awareness that time has passed, every belief change involves the replacement of old alternatives with new ones.[8] *Prima facie*, there's no reason to think that the beliefs of such subjects should be in any way constant—that their credences shouldn't be constantly ricochetting around simply due to the passage of time—without imposing a further constraint on their credences. What we need is a *Continuity Principle*, a principle that, in the appropriate circumstances, forces a subject's credences in new alternatives to be appropriately continuous with her credences in old alternatives. For subjects like us, virtually every diachronic argument with regards to what one's credences should be (including several that we'll look at in this paper) will require a principle of this kind to go through.

To cash out such a principle we need to answer two questions. First, what is it for a subject's credences in old and new alternatives to be 'appropriately continuous'? Second, what are the 'appropriate circumstances' in which a subject's credences in old and new alternatives should be continuous?

Let's start with the first question. Restrict our attention to the cases where the issue of continuity arises: belief changes where, at a given doxastic world, some old doxastic alternatives are eliminated and some new ones added. Say that an old and new alternative are *continuous* if a subject's credences in the old and new alternative should be 'appropriately continuous'.

The easiest case to consider is a belief change which just replaces one doxastic alternative at a world with another. If the new alternative is continuous with the old one, then it seems the subject should have the same credence in the new alternative as she had in the old alternative.

What about a belief change which just eliminates one alternative at a world and replaces it with two? Assume one of these new alternatives is continuous with the old

[8]This doxastic behavior also holds for some subjects who don't have a sense of time passing. An awareness of change and the knowledge that change requires the passage of time are sufficient.

one. It doesn't seem that the subject's credence in the new continuous alternative should need to be the same as her credence was in the old one. After all, her new epistemic situation is importantly different from her old one; she now has more alternatives at this world than she did before. I suggest that the intuitive relation between the subject's credences in the old and new continuous alternatives is the following: her credences should be such that if there was now a second belief change that just eliminated the other new alternative, then the subject's credence in the new continuous alternative should be the same as what her credence was in the old alternative.

What about a belief change which just removes two alternatives at a world and replaces it with one? Assume the new alternative is continuous with one of the old ones. Again, it doesn't seem that the subject's credence in the new alternative should need to be the same as her credence was in the old continuous alternative. After all, her new epistemic situation is importantly different from her old one; she now has fewer alternatives at this world than she did before. I suggest that the intuitive relation between the subject's credences in the old and new continuous alternatives is the following: the subject's credences should be such that if there was now a second belief change that just reintroduced the other old alternative, then the subject's credence in the new alternative should now be the same as her what credence was in the old continuous alternative.

More generally, the intuitive idea behind these cases is that if one alternative is continuous with another, it should be the case that in otherwise identical epistemic situations they should be alloted the same credence. Using this idea, we can provide a general characterization of what it is for a subject's credences in old and new alternatives to be continuous. Namely, if a belief change has a new alternative replace an old alternative, and the two alternatives are continuous, then a subject's credences should be such that if a second belief change reverted the subject's epistemic state back to how it was, with the exception of the new alternative taking the place of the old one, then her credence in the new alternative should be the same as what her credence was in the old one.

This characterization of continuity can be more simply captured if we assume centered conditionalization. On centered conditionalization, a sufficient and almost necessary condition for two alternatives to be continuous is that they have the same priors.

(The sole exception to this as a necessary condition is the trivial case in which a subject has only one doxastic alternative and has it replaced by another. In this case the old and new alternatives will be continuous regardless of her priors in them, since her credence in each alternative will trivially be 1.) So on centered conditionalization we can essentially think of continuous alternatives as alternatives with the same priors. For compartmentalized conditionalization this is not the case. While a pair of alternatives with the same priors will always be continuous, it will often be the case that two continuous alternatives will not have the same priors. (We'll return to the topic of compartmentalized conditionalization and continuity in section nine.)

Let's turn to the second question: what are the 'appropriate circumstances' in which a subject's credences in old and new alternatives should be continuous? I.e., when should an old and new alternative be continuous? This is a difficult question to answer. Intuitively, alternatives should be continuous when they're similar or related in the appropriate way. But it's hard to spell out what the right criteria are. I won't take a position here on what these standards should be. Instead, I'll take a Continuity Principle to be any principle which constraints rational credences such that some particular standard of continuity is preserved.

In the rest of this paper I'll assume that the subjects in question are like us, and have a sense of time passing. As a result, several of the arguments we'll look at in this paper will require a Continuity Principle of some kind to go through. In these places, I'll point out what standards of continuity are required. I won't take a position on whether these versions of the Continuity Principle are correct.

Given that subjects have a sense of time passing, it will often be convenient to leave the temporal shifting of alternatives implicit when describing belief changes, and to only explicitly mention the addition or elimination of alternatives not due to the passage of time. So, for example, consider a subject with two doxastic alternatives, one at which a coin toss comes up heads, the other at which the coin comes up tails. Suppose she learns the that the coin came up heads. If we leave temporal shifts implicit, we describe this belief change as eliminating her tails alternative. If we make temporal shifts explicit, we describe this belief change as eliminating both of her old alternatives, and adding a new tails alternative located at a later time. Leaving temporal shifting implicit allows us to

focus on the salient features of cases in which the passage of time is not the central issue, and allows us to concisely present cases where the temporal shifting of alternatives is straightforward.

4 Purely Self-Locating Belief Changes

I've claimed that changes in purely self-locating beliefs shouldn't affect our beliefs about what the world is like. Now I need to spell out what purely self-locating belief changes are.

In some cases it's uncontentious that changes in our self-locating beliefs do affect our *de dicto* beliefs. Consider a case where there appear to be two clocks in front of you. The one on the left reads 6 pm, and the one on the right reads 7 pm. You know that one of the clocks is in fact a ceramic sculpture that looks like a clock, while the other is a working clock that has been set to the right time. You don't what time it is, nor which clock is the sculpture. Now, if your self-locating beliefs were to change so that you believed it was 6 pm, you would change your beliefs about what the world is like: you would believe that in your world the clock on the right was a ceramic sculpture. Likewise, if you came to believe that in your world the clock on the right was a ceramic sculpture, you would come to believe it was 6 pm.

In cases like these self-locating beliefs are tied to *de dicto* beliefs. We want to separate cases like these from the cases that are contentious—cases in which I'll claim that self-locating belief changes shouldn't affect our *de dicto* beliefs.

A change in purely self-locating beliefs is a belief change which results in the addition or elimination of doxastic alternatives, but which doesn't result in the addition or elimination of doxastic worlds. So take the case of the two sages, where each is uncertain of who they are. If they come to believe that they are Zorn or Xingu, then they've eliminated the doxastic alternative where they are Xingu or Zorn. Their doxastic worlds are the same, though; they still believe they're in the world they're in. So this is a case of purely self-locating change.

Contrast purely self-locating belief changes with purely world-locating belief changes. A purely world-locating change is a belief change which adds or eliminates doxastic

worlds but does not otherwise affect the subject's doxastic alternatives. Examples of purely world-locating changes are hard to come by for subjects with a sense of time passing. Since for such subjects every belief change involves the replacement of old alternatives with new ones, virtually no belief change is purely world-locating.

Every belief change can be uniquely decomposed into a purely world-locating part and a purely self-locating part. The addition and elimination of doxastic alternatives at worlds that are added or eliminated is the purely world locating part of the belief change, and the addition or elimination of doxastic alternatives at worlds that aren't added or eliminated is the purely self-locating part of the belief change. For convenience, let 'purely self-locating change' apply both to belief changes that are purely self-locating and to the purely self-locating parts of belief changes. Likewise, let 'purely world-locating change' apply both to belief changes that are purely world-locating and to the purely world-locating parts of belief changes.

My claim is that in cases of purely self-locating change, one's credences in worlds, and thus in propositions, shouldn't change. This follows from *de dicto* conditionalization, and thus from compartmentalized conditionalization. On *de dicto* conditionalization a subject's credence in worlds is fixed by her doxastic worlds and her priors. Purely self-locating changes don't change a subject's doxastic worlds, and her priors are static. So purely self-locating changes won't affect one's credences in worlds.

For subjects like us, most purely self-locating changes are the boring purely self-locating changes brought about by our awareness of time passing. As we notice time pass we shift our alternatives, replacing each old alternative with a new one centered on the same individual but a later time. Exotic purely self-locating changes, where there are changes in the number of alternatives at a world, are rare. Many cases which seem to be exotic purely self-locating changes aren't purely self-locating changes at all.

As I'm writing this, I'm wondering what time it is. When I last looked at the clock it was 6 pm. Two of the times I think it might be are 7 pm and 7:05 pm. It might seem that looking at a clock and seeing that it's 7 pm is an exotic purely self-locating change, a change which eliminates the alternatives at my doxastic worlds which are located at the wrong time, but which leaves my doxastic worlds unchanged. 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 looking at the clock and finding out that it's 7 pm isn't an exotic purely self-locating change; time shifts aside, it's a purely world-locating change.[9]

We can see how little of our belief change is due to exotic purely self-locating changes in another way. 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 for 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.

For exotic purely self-locating change, one needs to increase or decrease the number of doxastic alternatives at a doxastic world. So there can only be cases of exotic purely self-locating change when there are doxastic worlds in which we have multiple doxastic alternatives, or in cases where we end up with multiple doxastic alternatives at a doxastic world. But doxastic worlds with multiple doxastic alternatives are strange worlds. They are worlds in which there are individuals-at-a-time that are in states subjectively

[9] A related case is given by Arntzenius (2003). A prisoner is put in a cell with two clocks, one that reads 6 pm and one that reads 7 pm. She knows that one of them is accurate, but not which, and her credences are evenly split between them. She further knows that her jailers will a flip a coin at midnight, and if it comes up heads, they'll turn off the lights. Arntzenius argues that if the lights are still on five hours from now, the prisoner's credences in head/tails should be $\frac{1}{3}/\frac{2}{3}$. What should the prisoner's credences be on centered and compartmentalized conditionalization?

Her credences are now evenly split between four sets of worlds: (1) worlds where the coin will be flipped six hours after she was put in the cell and the coin will come up heads, (2) worlds where the coin will be flipped five hours after she was put in the cell and the coin will come up heads, (3) worlds where the coin will be flipped six hours after she was put in the cell and the coin will come up tails, and (4) worlds where the coin will be flipped five hours after she was put in the cell and the coin will come up tails. If the lights are still on five hours from now, she'll eliminate the set of worlds where the coin will be flipped five hours after she was put in the cell and the coin came up heads. This is a purely world-locating change, not a purely self-locating change, so centered and compartmentalized conditionalization will treat it the same way: her credence in each of the remaining sets of worlds will increase to $\frac{1}{3}$, so her credence in heads/tails will be $\frac{1}{3}/\frac{2}{3}$.

indistinguishable from other individuals-at-a-time.

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. At the worlds I think are likely, exotic purely self-locating changes don't arise. Only at rare fringe worlds are there exotic purely self-locating changes, and my credence in these worlds is so small that these changes have little affect on my overall belief distribution.

5 Sleeping Beauty

We saw a case of exotic purely self-locating change above, with the two sages. If the two sages come to believe that they're Zorn, they've gone from having two doxastic alternatives at their world to one. A more interesting case of exotic purely self-locating 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 the coin comes up heads, they will wake you up once during that time; if it comes up tails, they will wake you up twice. 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 awakening.

When you wake up, what should your credence be that the coin came up heads? If you then learn that it's Monday, what should your credence in heads become?

On Sunday you have $\frac{1}{2}$ credence that you're in a world where the coin will come up heads, and a $\frac{1}{2}$ credence that you're in a world where the coin will come up tails. Assume that on Sunday you have one doxastic alternative at each of these doxastic worlds. When you wake up this is no longer the case. At each of the tails worlds you

now have two doxastic alternatives—one where it's Monday and one where it's Tuesday. This is a purely self-locating change, so on compartmentalized conditionalization your credence in the proposition that heads came up should remain $\frac{1}{2}$. Likewise for tails.

Given tails, what should your credence be that it's Monday versus Tuesday? Your credence in the tails worlds is $\frac{1}{2}$, so your credences in these two alternatives must sum to $\frac{1}{2}$. By compartmentalized conditionalization, the $\frac{1}{2}$ should be divided such that the ratio between the two alternatives is the same as the ratio between your hypothetical priors in these alternatives.

Those swayed by Indifference Principles will advocate assigning equal credences to these two alternatives. Indifference Principles can be seen as rationality constraints on one's priors. An Indifference Principle requires that one's priors be such that whenever one is 'indifferent' between several possibilities (by some standard of indifference), one's credences in these possibilities are equal. I generally don't find Indifference Principles very compelling, but they're a convenient way to generate examples of permissible belief distributions. So for convenience I'll use something like an Indifference Principle as a heuristic by which to assign specific values.

So on the account I favor, when you wake up your credences in heads and tails should be $\frac{1}{2}/\frac{1}{2}$, with your credence in tails split evenly between Monday and Tuesday. How should your credences change if you then learn that it's Monday? This information eliminates a doxastic alternative at the tails worlds, the alternative where it's Tuesday. This is another purely self-locating change, so your credences in heads and tails should remain $\frac{1}{2}/\frac{1}{2}$.

Two other responses to the sleeping beauty problem have been advocated in the literature. The majority of the literature on sleeping beauty has endorsed the response offered by Adam Elga (2000).[10] Elga proposes that upon awaking 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 then learn that it's Monday, you conditionalize and regain your original $\frac{1}{2}/\frac{1}{2}$ credences in heads and tails.

[10]See Dorr (2002), Monton (2002), Arntzenius (2003), and Hitchcock (2004).

The other response is Lewis' (2001). Lewis proposes that we retain our $\frac{1}{2}/\frac{1}{2}$ credences in heads and tails when we wake up, with our credence in tails split evenly between Monday and Tuesday. Lewis' account diverges from the account I favor in what happens when you then learn that it's Monday. Lewis holds that you should conditionalize and come to have a $\frac{2}{3}$ credence in heads and a $\frac{1}{3}$ credence in tails.

We can see the disagreements between the account I favor and their accounts in terms of how we think changes in purely self-locating beliefs affect our *de dicto* beliefs. On the account I favor, our *de dicto* beliefs are held fixed in all cases of purely self-locating change. So neither waking up in the sleeping beauty case nor then being told it's Monday changes your credences in heads and tails.

On Lewis' account increasing the number of doxastic alternatives at a world doesn't alter our *de dicto* beliefs, but decreasing the number of doxastic alternatives does; namely, decreasing the number of doxastic alternatives at a world decreases our credence in that world. So waking up in the sleeping beauty case doesn't change our credences, but then being told it's Monday does, decreasing our credence in tails.

On Elga's account both kinds of purely self-locating changes affect our *de dicto* beliefs. Increasing the number of doxastic alternatives at a world increases our credence in that world, and decreasing the number of doxastic alternatives at a world decreases our credence in that world. So waking up in the sleeping beauty case increases our credence in tails, and then being told it's Monday decreases our credence in tails.

We can also see the differences between these accounts in terms of which generalization of conditionalization one endorses with regards to *de se* beliefs. The response to the sleeping beauty case I've offered above follows straightforwardly from compartmentalized conditionalization and some minimal assumptions about priors. We'll see that Elga's and Lewis' responses rely on centered conditionalization.

In the next section I'll analyze an argument for Elga's response, and show that Elga's response leads to some highly counterintuitive results. (A different kind of argument for Elga's response using dutch books doesn't neatly intersect with the issues I look at here. As it turns out, dutch books do little to settle the issue between centered and compartmentalized conditionalization. Since these issues are given a satisfactory treatment in

Halpern (2004), I'll restrict my comments on them to the following footnote.[11]) I'll then look at Lewis' position. I'll show that while Lewis' account escapes the difficulties facing Elga's response, it ends up facing other, equally serious, difficulties.

A caveat: I present Elga's and Lewis' arguments in my own terms. In some places I've filled in and spelled out implicit premises that the arguments require. I believe the arguments I present are faithful to the original arguments, but it matters little for my purposes if they're not. The arguments I do present will lead us to interesting results all the same.

6 Elga's Response to Sleeping Beauty

The sleeping beauty case is an instance of a general type of case. The same questions arise for cases involving duplication or fission. Consider a case where instead of waking you up twice if the coin comes up tails, the researchers create a duplicate of you in a distant, qualitatively identical location. On Sunday your credences in heads and tails are $\frac{1}{2}/\frac{1}{2}$. What should your credences in heads and tails be when you wake up on Monday? Given tails, what should your credence be that you're the duplicate? I'll take it that the answers given for the sleeping beauty case apply here as well. This needn't be the case, of course. One might try to treat sleeping beauty-type cases differently from duplications cases, duplication cases differently from fission cases, and so on. But for the purposes

[11]The cleanest presentation of a dutch book argument for Elga's response is given in Hitchcock (2004). Hitchcock argues that the salient dutch book in the sleeping beauty case is one in which we should adopt 2:1 odds on tails. The proponent of compartmentalized conditionalization will agree that in the dutch book Hitchcock describes, one should accept 2:1 odds on tails as fair. But she will disagree that this suggests her credences should be $\frac{1}{3}/\frac{2}{3}$. This is because she will consider it to be a case of double counting—tails payoffs are enacted twice, while heads payoffs are enacted only once—so someone with equal credences in H/T should accept 2:1 odds. What the dutch book suggets, she will argue, is that her credences in H/T should be $\frac{1}{2}/\frac{1}{2}$. After all, if she had $\frac{1}{3}/\frac{2}{3}$ credences in H/T, and tails payoffs are counted twice, she should be accepting 4:1 odds, not the 2:1 odds the dutch book suggests.

The fallout, I think, is that dutch books do little to settle the issue between centered and compartmentalized conditionalization. While both positions will generally agree on how one should bet, they will disagree on what implications this has with regards to one's credences. These issues are addressed in Halpern (2004). The response to the dutch book argument given above first appears in Arntzenius (2002), along with some interesting thoughts about the relation between these issues and decision theory.

of this paper, I'll assume that they should be treated the same way.

Elga's argument for his response follows from four principles:

1. An Indifference Principle
2. A Continuity Principle
3. Centered Conditionalization
4. The Principal Principle

Let $D(\cdot)$ be your credence function, H/T be the propositions that the coin came up heads/tails, and MON/TUE the propositions that it is Monday/Tuesday. The first step of Elga's argument uses an Indifference Principle to argue that upon awakening $D(T\wedge MON) = D(T\wedge TUE)$. The second step of the argument uses a Continuity Principle, the Principal Principle and centered conditionalization to argue that upon awakening $D(H\wedge MON) = D(T\wedge MON)$. From these two steps it follows that $D(H\wedge MON) = D(T\wedge MON) = D(T\wedge TUE) = \frac{1}{3}$. The third step of the argument uses the Continuity Principle and centered conditionalization to argue that if you learn it's Monday after awaking, then $D(H\wedge MON) = D(T\wedge MON) = \frac{1}{2}$.

The first step of the argument uses a restricted version of the Indifference Principle, proposed and defended by Elga (2004). According to Elga's principle you should have the same credences in subjectively identical doxastic alternatives at the same doxastic world. Since I'm assuming transparency—that subjects always have access to their own subjective states—a subject's doxastic alternatives are always subjectively identical, and satisfying this principle entails having the same credences in all doxastic alternatives at the same doxastic world. After you wake up in the sleeping beauty case you have two doxastic alternatives at each tails world: $T\wedge MON$ and $T\wedge TUE$. By this principle $D(T\wedge MON) = D(T\wedge TUE)$.

The second and third steps of the argument use a Continuity Principle. As we've seen, the content of such a principle depends on the standards of continuity employed. For Elga's argument, any Continuity Principle for which the following is a sufficient condition for continuity will do. Consider a belief change which eliminates some old alternatives at a world and replaces them with new ones. An old alternative and new

alternative should be continuous if: (a) both alternatives are focused on the same person p, (b) of the new alternatives focused on p, this new alternative is the located at the earliest time following the time of the old alternative.

Now consider the belief change that takes place between going to sleep on Sunday and waking up on Monday in the sleeping beauty case. Assume our belief changes are governed by centered conditionalization, and that the purely world-locating part of this change, if any, won't affect our credences in heads and tails. (For conciseness, I'll leave the role that centered conditionalization plays in this argument—connecting up our credences and priors—implicit.) By the Principle Principal, our original credences in H∧SUN and T∧SUN are $\frac{1}{2}/\frac{1}{2}$, and thus our priors in H∧SUN and T∧SUN are equal. From the Continuity Principle it follows that H∧SUN and T∧SUN are continuous with H∧MON and T∧MON, respectively, and thus that our priors in H∧SUN and T∧SUN are the same as our priors in H∧MON and T∧MON. So our priors in H∧MON and T∧MON are equal, and thus so are our credences.

So the Continuity Principle, centered conditionalization and the Principal Principle entail that D(H∧MON) = D(T∧MON), and the restricted Indifference Principle entails that D(T∧MON) = D(T∧TUE). Combining these results, it follows that upon awaking in the sleeping beauty case one's credences should be D(H∧MON) = D(T∧MON) = D(T∧TUE) = $\frac{1}{3}$.

Say you're woken up at 9 am. What if at 9:01 am you learn that it's Monday? Take the belief change between 9 and 9:01 am, and assume that the purely world-locating part of this change won't change our credences in heads and tails. From the Continuity Principle it follows that H∧MON(9 am) and T∧MON(9 am) are continuous with H∧MON(9:01 am) and T∧MON(9:01 am), respectively, and thus that our priors in H∧MON(9 am) and T∧MON(9 am) are the same as our priors in H∧MON(9:01 am) and T∧MON(9:01 am). We saw above that our priors in H∧MON(9 am) and T∧MON(9 am) are equal, so our priors in H∧MON(9:01 am) and T∧MON(9:01 am) must be equal as well. So our credences after being told it's Monday should be D(H∧MON(9:01 am)) = D(T∧MON(9:01 am)) = $\frac{1}{2}$.

Note that the Principal Principle only plays a superficial role in Elga's argument. The Principal Principle sets our credences in heads and tails 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.[12] 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 awaking 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. 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, if we conclude that we don't receive inadmissible evidence upon awaking we still need to decide what to say about Elga's argument, 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 wake up.) 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, then purely self-locating changes that increase the number of doxastic alternatives at a world will increase one's credence in that world relative to worlds without such an increase. Likewise, one's credence in a proposition

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

which multiplies doxastic alternatives will 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 leaves one open to the following objection, based on an argument pointed out to me by Tim Maudlin:

> *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, 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 one. 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.

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 the vat.

The many brains argument assumed that brain in the vat duplication is the only proposition you have a non-zero credence in 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 worlds. 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 world. 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. In section four I argued that worlds

with multiple doxastic alternatives 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.

7 An Escape Route?

To escape these many brains-type arguments we need to reject Elga's argument. Elga's argument relied on four principles: the Principal Principle, an Indifference Principle, a Continuity Principle and centered conditionalization.

What if we reject the Principal Principle? We saw above that the Principal Principle only plays a superficial role in the Elga's argument; any means of assigning $\frac{1}{2}/\frac{1}{2}$ credences to heads/tails will do. In the many brains-type arguments the Principal Principle plays no role at all; these arguments go through given any assignment of non-zero credences to the relevant propositions.

Well, what if we do assign a 0 credence to the relevant propositions, those that multiply our doxastic alternatives? That is, what if we have 0 priors in any world that multiplies our doxastic alternatives?[13] It will still be true that purely self-locating changes that increase the number of doxastic alternatives at a world will increase one's credence in that world, but it will be true trivially, since there will be no purely self-locating changes which increase the number of doxastic alternatives at a world. Any world where this might have happened gets a 0 prior.

This is counterintuitive, since we can imagine cases where it seems we have very good evidence that we're in a world where our doxastic alternatives are being multiplied. Consider a scientist who has invented brain-in-the-vat duplication technology (though she's never tried this on herself), and who has just turned on a machine that creates brains in states subjectively identical to her own. On this approach, such a scientist should have a 0 credence in the machine working and successfully creating epistemic duplicates of herself, even if she has no reason to think anything will interfere and

[13]The same purpose could be achieved by assigning an infinitesimal credence to such worlds, as long as worlds like our own were still assigned finite credences. Assigning infinitesimal credences raises the same problems as assigning 0 credences.

believes the machine to be in perfect working order!

I don't think this response to the sleeping beauty case is very satisfying. In any case, *my* priors in worlds that multiply alternatives aren't all 0, I don't accept the conclusion of the many brains-type arguments, and I don't feel particularly irrational (in this regard). So I'm inclined to look for a different solution.

What if we reject Elga's Indifference Principle? As with the Principal Principle, simply rejecting the Indifference Principle isn't enough to escape the many brains-type arguments. What is needed is for the sum of our credences in the alternatives at an alternative multiplying world to converge to a value less than one.

Given the Continuity Principle and centered conditionalization, new alternatives will get the same priors as the alternatives they're continuous with. So we can't play with the priors assigned to continuous alternatives. To get the sum of our credences in the alternatives at alternative multiplying worlds to converge to less than one, we need to restrict the priors assigned to discontinuous alternatives. The simplest way to do this is to have a 0 prior in every centered world that would otherwise become a discontinuous alternative. A less extreme option is to have priors such that finite but decreasing credences are assigned to new discontinuous alternatives, such that the sum of our credences in the alternatives at that world converge to some value less than one.

I said earlier that I generally don't find Indifference Principles very compelling as a constraint on rational priors. At the same time, it's plausible that we often do have roughly equal credences in possibilities we're in some sense indifferent between, such as doxastic alternatives at the same world. I also think that most of us think we don't live in a strange world, and aren't irrational in thinking this. If I'm right, then it's not plausible to insist that having an 'indifferent' distribution commits us to thinking we live in a strange world, and we should look for a different way to avoid the many brains-type arguments.

We're left with two ways to escape the many brains-type arguments: we can deny Elga's Continuity Principle or we can deny centered conditionalization. It's not surprising that these are the two remaining options. The source of the difficulty is that if one accepts Elga's argument, purely self-locating changes that increase the number of doxastic alternatives at a world will increase one's credence in that world (relative to worlds

without such an increase). But the Continuity Principle and centered conditionalization alone entail this.

Consider two doxastic worlds, A and B. At t_1 we have n alternatives at each world. At t_2 a purely self-locating change adds m alternatives to A. By the Continuity Principle and centered conditionalization, our priors in n of the alternatives at each world at t_2 will be the same as our priors in their predecessors. But world A gets m additional alternatives with non-zero priors. So our credence in A will increase relative to our credence in B.

Given the Continuity Principle and centered conditionalization, some further auxiliary assumptions—that one has non-zero credences in alternative multiplying hypotheses, etc.—are still needed to get the many brains-type arguments to work. But the Continuity Principle and centered conditionalization are the crucial elements. A plausible response to the many brains-type arguments is going to need to reject one of them.

8 Lewis' Response to Sleeping Beauty

Lewis' discussion of the sleeping beauty case questions Elga's response by looking at the Principal Principle and at whether any inadmissible evidence is received upon awaking. Lewis contends that no evidence is received, and thus that the Principal Principle should still apply. 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. So how would Lewis address Elga's argument? Lewis accepts a Principal Principle that entails that our credences in heads and tails on Sunday should be $\frac{1}{2}/\frac{1}{2}$. Furthermore, Lewis explicitly endorses (centered) conditionalization and the Indifference Principle. So he must reject Elga's Continuity Principle.

Lewis' argument requries a somewhat different Continuity Principle, one compatible with the following two conditions. First, an old alternative and new alternative at a world should be continuous if: (a) both alternatives are focused on the same person p, (b) of the new alternatives focused on p, this new alternative is the located at the earliest time following the time of the old alternative, (c) the belief change hasn't increased the number of alternatives at that world. Second, an old alternative and new alternative

should never be continuous if both (a) and (b) hold and the belief change has increased the number of alternatives at that world.

So how are credences assigned in cases where the number of alternatives at a world increases? Lewis' position seems to be that in cases of purely self-locating change where the number of alternatives increases at a world, we get no evidence with regards to what world we're in. Call this the Increasing No-Evidence Principle: in cases of purely self-locating change where the number of alternatives at a world increases, our credence in that world should remain the same.

So Lewis' argument for his response uses five principles:

1. An Indifference Principle
2. A Continuity Principle
3. Centered Conditionalization
4. The Principal Principle
5. The Increasing No-Evidence Principle

By the Principal Principle our credences in heads and tails on Sunday should be $\frac{1}{2}/\frac{1}{2}$. As before, assume that the purely world-locating part of the belief change between Sunday and Monday (if any) won't affect our credences in heads and tails. The purely self-locating part of this change increases the number of alternatives at tails worlds, so by the Increasing No-Evidence Principle our credence in tails should remain the same on Monday as it was on Sunday: $\frac{1}{2}$. So our credence in heads on Monday should be $\frac{1}{2}$ as well. By the Indifference Principle, our credences in T$\wedge$MON and T$\wedge$TUE should be the same, so D(T$\wedge$MON) = D(T$\wedge$TUE) = $\frac{1}{4}$ and D(H$\wedge$MON) = $\frac{1}{2}$.

What if you then learn at 9:01 am that it's Monday? The reasoning here is the same as before. Take the belief change between 9 and 9:01 am, and assume that the purely world-locating part of this change won't change our credences in heads and tails. By the Continuity Principle and centered conditionalization, our priors in H$\wedge$MON(9 am) and T$\wedge$MON(9 am) should be the same as our priors in H$\wedge$MON(9:01 am) and T$\wedge$MON(9:01 am), respectively. Our prior in H$\wedge$MON(9 am) is twice that of our prior in T$\wedge$MON(9 am), so our prior in H$\wedge$MON(9:01 am) should be twice that of our prior

in T∧MON(9:01 am). So our credences after being told it's Monday should then be D(H∧MON(9:01 am)) = $\frac{2}{3}$, D(T∧MON(9:01 am)) = $\frac{1}{3}$.

Elga's argument ran into problems because it entailed that purely self-locating changes that increased the number of alternatives at a world increased one's credences in that world, relative to worlds without such an increase. Lewis avoids this result by rejecting Elga's Continuity Principle and adopting the Increasing No-Evidence Principle. However, if one accepts either Elga or Lewis' argument, then purely self-locating changes that decrease the number of doxastic alternatives at a world will decrease one's credence in that world, relative to worlds without such a decrease. And this also leads to counterintuitive consequences. Namely, accepting Lewis' argument leaves one open to the following objection:

> *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 (Increasing No-Evidence Principle), and this credence will be evenly split between your $n+1$ alternatives (Indifference Principle). As a half second second passes and these brains are destroyed, your credence that you are in such a world will decrease by the appropriate amount (Lewis' Continuity Principle and centered conditionalization). So as each second passes, your credence that you are in such a world will decrease. 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.

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-the-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 account I am attributing to Lewis, then though you're now almost certain that she will carry out her threat, you'll be almost certain that she didn't when you wake up tomorrow morning. 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!

The difficulty stems from the fact that purely self-locating changes that decrease the number of alternatives at a world decrease one's credence in that world (relative to worlds without such a decrease). This follows directly from centered conditionalization and either Lewis' or Elga's Continuity Principles. Take two doxastic worlds, A and B. At t_1 we have n alternatives at each world. At t_2 a purely self-locating change eliminates m of the alternatives at A (where $m < n$). By centered conditionalization and either Continuity Principle, our priors in n of the alternatives at A and $n-m$ of the alternatives at B at t_2 will be the same as our priors in their predecessors. But world A loses m alternatives, so our credence in A will decrease relative to our credence in B.

As with the many brains-type arguments, several further assumptions are needed for the sadistic scientists-type arguments to go through. But the Continuity Principle and centered conditionalization are the crucial elements. We saw that a plausible rejection of the many brains-type arguments required that we reject either centered conditionalization or Elga's Continuity Principle. To escape the sadistic scientists-type arguments as well our choices are further restricted: we must reject either centered conditionalization or both Elga's and Lewis' Continuity Principles.

9 Further Considerations

I've presented three positions on the dynamics of *de se* beliefs: Elga's stance, Lewis' stance, and the stance of Halpern, Tuttle and I. In the previous three sections I've argued that Elga's and Lewis' positions lead to highly counterintuitive consequences. In this section I'll consider some further reasons for favoring some of these stances over the others, and look to see if a skeptical scenario can be raised against compartmentalized conditionalization as well.

9.1 Continuity

In section three we saw that we need a continuity principle to keep our credences from varying wildly as time passes. As with any constraint on credences in a Bayesian framework, this principle can be reformulated as a constraint on our priors—as a principle which only allows priors such that a subject with those priors who updates properly will always have credences that satisfy the original constraint. How severe a constraint on our priors does a continuity principle impose?

In section four we saw that almost all of our purely self-locating changes are boring purely self-locating changes, where each old alternative at a world is replaced with a new one centered on the same individual but a later time. Likewise, we saw that at almost all of our doxastic worlds we only have one doxastic alternative. Restrict our attention to these cases: boring purely self-locating changes at worlds with one alternative. Continuity requires a hefty constraint on our priors if we adopt centered conditionalization, in both these cases and the general case. What if we adopt compartmentalized conditionalization?

If we adopt compartmentalized conditionalization, continuity is free in these cases. Compartmentalized conditionalization divides the credence allocated to a world among the alternatives at that world in accordance with their priors. But if there's only a single alternative at a world, then it will be assigned all of the credence allocated to that world, regardless of its prior. So in the vast majority of cases we deal with—boring changes at worlds with one alternative—continuity falls right out of the dynamics! This is a mark in favor of compartmentalized conditionalization.

9.2 Reflection

Van Fraassen (1984) has suggested that we adopt a Reflection Principle as a constraint on rationality. The Reflection Principle is, roughly, that your conditional credence in h given that your credence in h will become x, should be x.[14] In the special case where you believe that your credence in h will be x, reflection entails that your current credence in h should be x.

As with conditionalization, one can consider *de dicto* and *de se* versions of reflection, with h ranging over *de dicto* and *de se* propositions, respectively. Unlike conditionalization, the *de se* version of reflection is not very interesting. The *de se* version of reflection is untenable, and there's no straightforward way of fixing it up.[15] Consider the centered proposition that it's 7 p.m. The *de se* version of reflection requires that if you believe that at some time in the future you'll believe it's 7 p.m., you should believe it's 7 p.m now. But it can be rational to both believe that it's 6 p.m. now and believe that at some time in the future you'll believe it's 7 p.m. You might, for example, be in possession of a reliable watch. From now on, let us restrict our attention to the *de dicto* version of reflection.

Unrevised *de dicto* conditionalization along with some further assumptions, such as that one's potential evidence forms a partition of possible worlds, entails reflection.[16] Revised *de dicto* conditionalization violates reflection, since it allows evidential instances to add doxastic worlds; i.e., allows a subject to lose information about the world.[17]

[14] I say "roughly" because this version of the principle is narrower than the principle of reflection Van Fraassen (1995) now subscribes to. Van Fraassen now subscribes to a principle he calls General Reflection, from which the above principle can be derived as a special case.

[15] Intuitively, what one wants is a tie between what one thinks their future credences in a centered proposition will be, and one's current belief in some suitable correlate to that centered proposition. But it's hard to see what such a 'suitable correlate' would be. It seems, for example, that such a correlate would not be representable as any kind of belief in the standard Lewisian framework (as a set of centered worlds).

[16] This is the standard claim made about the relationship between conditionalization and reflection. Weisberg (2004) argues that this is not the case; what the entailment requires is not that the subject conditionalize, but that she *believes* she conditionalizes. Likewise, the entailment requires that the subject *believes* her potential evidence forms a partition. For simplicity, I'll continue to write in this section as if the standard claim were correct; it is a simple matter of introducing the appropriate 'if she believes' clauses to bring the discussion here in line with Weisberg's claim.

[17] Again, see Arntzenius' (2003) Shangri-la example for a case in which this happens.

Centered and compartmentalized conditionalization also violate reflection for this reason.

However, centered conditionalization can violate reflection even if we exclude belief changes that add doxastic worlds. This is not very surprising; *de dicto* conditionalization is closely tied to reflection, and centered conditionalization and *de dicto* conditionalization conflict. We've already seen an example of such a violation in the Sleeping Beauty case. Given centered conditionalization, your Sunday credences that the coin will come up heads/tails is $\frac{1}{2}/\frac{1}{2}$, even though you know that your credences on Monday in heads/tails will be $\frac{1}{3}/\frac{2}{3}$.

Compartmentalized conditionalization avoids some of the reflection violations of centered conditionalization. On compartmentalized conditionalization, potential purely self-locating changes that add or eliminate doxastic alternatives don't bring about violations of reflection.[18] So in the sleeping beauty case your credence in the proposition that heads/tails will come up does not violate reflection: your credence in heads/tails on Sunday is $\frac{1}{2}/\frac{1}{2}$, and you know your credence in heads/tails on Monday will still be $\frac{1}{2}/\frac{1}{2}$.

But compartmentalized conditionalization also violates reflection in some ways that centered conditionalization does not. This is surprising since *de dicto* conditionalization is built into compartmentalized conditionalization, and, given some apparently innocuous assumptions, *de dicto* conditionalization entails reflection. Compartmentalized conditionalization doesn't entail reflection because one of the assumptions required for the entailment fails when irreducibly *de se* evidence is taken into account. Namely, in cases where evidence can be irreducibly *de se*, the potential evidence need not form a partition of possible worlds.

Take a case like the sleeping beauty case, but with the following twist: if the original coin toss comes up tails, they'll put you in a black room on Monday and a white room on Tuesday. If the original coin toss comes up heads, another coin will be flipped to determine whether to put you in a black room or a white room on Monday. In this case one's potential evidence is either waking up and seeing a black room or waking

[18]Halpern (2004) considers some further issues regarding reflection in the case of subjects who, unlike the subjects we're considering here, need not have a sense of time passing. He shows that an agent who satisfies a condition he calls *perfect recall* and updates by compartmentalized conditionalization will satisfy a version of reflection, while an agent who satisfies perfect recall but updates by centered conditionalization will not.

up and seeing a white room. These two pieces of evidence don't form a partition of possible worlds because they're not mutually exclusive with regards to worlds. Both pieces of evidence are compatible with the worlds where the original coin toss comes up tails: seeing black with the tails and Monday alternative, seeing white with the tails and Tuesday alternative.

In this case compartmentalized conditionalization violates reflection: your credences in heads and tails on Sunday will be $\frac{1}{2}/\frac{1}{2}$, but you know that on Monday after you open your eyes your credences in heads and tails will be $\frac{1}{3}/\frac{2}{3}$. If you see a black room, you'll eliminate half of your heads worlds—the worlds where the second coin toss came up such that they put you in a white room. You'll also eliminate the Tuesday alternatives at your tails worlds, but this won't eliminate any of the tails worlds. So on compartmentalized conditionalization, if you see a black room your credence in tails will go up. Likewise, if you see a white room, you'll eliminate half of your heads worlds, and eliminate the Monday alternatives at your tails worlds. But none of the tails worlds will be eliminated, so again your credence in tails will go up.

Centered conditionalization will also violate reflection in this case, of course, for the same reason it violates reflection in the standard sleeping beauty case. On centered conditionalization, though, this violation will take place between Sunday and when you wake up on Monday, and no further violation will take place between Monday before you open your eyes and Monday after you open your eyes. On compartmentalized conditionalization, the opposite is the case: no reflection violation takes place between Sunday and Monday before you open your eyes, but a reflection violation does take place between Monday before you open your eyes and Monday after you open your eyes.

Elga's position has struck many as strange because it allows changes in a subject's *de dicto* beliefs without, intuitively, the subject having gained or lost any information about what the world is like. In the black and white room case, for example, your credence in heads and tails changes *before* you open eyes, even though, intuitively, you have the same information about what the world is like as you had before you went to sleep. On the account I favor there is no such oddity. In the black and white room case, there's no change in your *de dicto* beliefs until you open your eyes and see that you're in (say) a black room. And it's this *de dicto* information—that you're not in a heads and

white room world—that directly brings about the change in your *de dicto* credences. If there's no *de dicto* information to be had—as in the sleeping beauty case—then there's no change in your *de dicto* credences, and no violations of reflection. I take this to be a mark in favor of compartmentalized conditionalization.

I've focused on Elga's account, but Lewis' account has the same deficits. While Lewis does not allow the addition of doxastic alternatives to change our beliefs about what the world is like, he does allow the elimination of doxastic alternatives to change our beliefs about what the world is like. Thus he is open to the same criticisms: he allows changes in a subject's *de dicto* beliefs without, intuitively, the subject having gained or lost any information about what the world is like.

9.3 The Varied Brains Argument

Reflection considerations aside, the black and white room case raises a natural worry for the account I favor. 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 the set of doxastic worlds remains the same, our credences in the respective 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:

> *The Varied Brains Argument:* Assume your credences are divided between two kinds of worlds, normal (N) worlds and strange (S) worlds. Among all these worlds there are n subjectively distinguishable experiences, E_1 through E_n, that you might experience in the next second. In each of your doxastic S-worlds scientists are creating n brains in vats in the following second, each one compatible with some E_i.

> In your doxastic N-worlds you have no subjective duplicates, but you have some N-world compatible with each E_i. Now, at the end of a second you'll have experienced some E, say E_1. This will eliminate the doxastic N-worlds incompatible with E_1, i.e., the ones that were compatible with E_2 through E_n. On the other hand, all of your doxastic S-worlds are compatible with E_1, so no doxastic S-world will be eliminated. By compartmentalized conditionalization, your credence in the S-worlds relative to the N-worlds will increase.
>
> We can extend this case by replacing 'second' with longer and longer units of time, and as the unit of time grows larger, the number n of distinguishable experiences you might experience during this period likewise grows larger. By making the unit of time arbitrarily large, we can get a case in which on compartmentalized conditionalization one's credence in the S-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, D(S) gains on D(N) because of the artificial way in which the doxastic worlds have been selected: all the strange worlds under consideration are ones that will end up matching what we experience, whereas many of the normal worlds that are considered won't end up matching what we experience. If we restricted the normal worlds to those compatible with E_1, D(S) wouldn't gain on D(N). Likewise, if we placed no restrictions on which strange worlds were allowed, then E_1 would eliminate lots of strange worlds as well as lots of normal worlds. Whether D(S) gains on D(N) depends on which S and N-worlds are doxastic worlds—which worlds our priors and evidence lead us to believe could be ours. And it's reasonable to think that D(S) will not gain on D(N) for people with doxastic worlds like ours.

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 highly 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.

10 A Choice

The counterintuitive aspects of Elga's and Lewis' responses can be brought together into a single case:

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

If you endorse Elga's solution to the sleeping beauty case, 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' solution to the sleeping beauty case, then your credence that the coin came up heads will converge to 1, again regardless of your evidence (knowledge of objective chances, etc.) to the contrary.

I take both of these outcomes to be counterintuitive. Intuitively, our credences should remain $\frac{1}{2}/\frac{1}{2}$ throughout. Intuitively, purely self-locating changes don't provide us with any new information about the world, and shouldn't change our credences about what the world is like. Intuitively, purely self-locating changes shouldn't change our credences in propositions. Call this CLAIM.

There are two ways to satisfy CLAIM. First, we can reject centered conditionalization and adopt a different belief dynamics, preferably one more compatible with CLAIM. One choice for such a dynamics is compartmentalized conditionalization, which satisfies CLAIM automatically.

Once we've entertained the notion of compartmentalizing beliefs, however, a number of possibilities arise, such as belief dynamics with several layers of compartmentalization, belief dynamics that compartmentalize groups of worlds or within worlds, etc. Compartmentalized conditionalization is uniquely picked out if we add two further constraints. First, require that the dynamics be compatible with *de dicto* conditionalization. This entails that the dynamics must first compartmentalize at the level of worlds. Second, require that the dynamics treat sleeping beauty cases the same way as duplication cases, duplication cases the same way as fission cases, and so on. That is, require that the dynamics treat shifts in the centered worlds at a world uniformly, taking into consideration only the change in numbers of centered worlds and the relevant priors, not the features of the centered worlds. This eliminates any dynamics that compartmentalize non-trivially within worlds. Only compartmentalized conditionalization meets these two constraints. Neither of these constraints is beyond question, but I take them both to be *prima facie* plausible. So I take compartmentalized conditionalization to be a natural choice for the dynamics.

Second, we can keep centered conditionalization and constrain our priors such that our belief changes will be compatible with CLAIM.[19] This includes rejecting the Con-

[19]This is only non-trivially possible for agents who have a sense of time passing. For these agents, priors in alternatives can be contrived such that CLAIM is non-trivially satisfied.

tinuity Principles which Elga's and Lewis' arguments require to go through. But just rejecting these principles isn't enough. This option requires adopting a strengthened version of the Increasing No-Evidence Principle: in cases of purely self-locating change where the number of alternatives at a world increases, decreases, or stays the same, our credence in that world should remain the same. This general No-Evidence Principle entails CLAIM.

Which of the two options should we choose? I suggest that we choose the first option and adopt compartmentalized conditionalization. If we adopt compartmentalized conditionalization we get CLAIM for free. If we adopt centered conditionalization we only get CLAIM after imposing draconian restrictions on our priors, restrictions that in effect make our belief changes look like they're being governed by compartmentalized conditionalization. What's the point of adopting centered conditionalization if what we really want is for our beliefs to behave as if we'd adopted compartmentalized conditionalization?[20]

[20]I'd like to thank Frank Arntzenius, Maya Eddon, Adam Elga, Hilary Greaves, John Hawthorne, David Manley, Tim Maudlin, Adam Sennet and Jonathon Weisberg for valuable comments and discussion. In particular, I owe much to David Manley, for raising the black and white room case, and to Tim Maudlin, who's many worlds argument inspired my interest in these issues. Finally, I owe a special thanks to Frank Arntzenius for comments on a number of drafts, and endless barroom discussion. The bulk of this work was completed with gracious funding from Rutgers University in the Fall of 2003.

References

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

Arntzenius, F. (2003) "Self-locating Beliefs, Reflection, Conditionalization and Dutch Books", *Journal of Philosophy*, pp. 356-70

Dorr, C. (2002) "Sleeping Beauty: in defense of Elga", *Analysis*, pp. 292-6

Earman, J. (1986) *Bayes or Bust?*

Elga, A. (2000) "Self-locating belief and the Sleeping Beauty problem", *Analysis*, pp. 143-7

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

Halpern, J. and Tuttle, M. (1993) "Knowledge, probability, and adversaries", *Journal of the ACM*, pp. 917-63

Halpern, J. (2004) "Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems", *Proceedings of the Twentieth Conference on Uncertainty in AI*, pp. 226-34

Hitchcock, C. (2004) "Beauty and the Bets", *Synthese*, pp. 405-20

Lewis, D. (1979) "Attitudes *De Dicto* and *De Se*" in *The Philosophical Review*, pp. 513-43

Lewis, D. (1980) "A Subjectivist's Guide to Objective Chance" in *Studies in Inductive Logic and Probability, Vol. 2*

Lewis, D. (1983) "Individuation by acquaintance and by stipulation" in *The Philosophical Review*, pp. 3-33

Lewis, D. (2001) "Sleeping Beauty: reply to Elga", *Analysis*, pp. 171-6

Monton, B. (2002) "Sleeping Beauty and the forgetful Bayesian", *Analysis*, pp. 47-53

Strevens, M. (2004) "Bayesian confirmation theory: Inductive logic or mere inductive framework?", *Synthese*, forthcoming

Van Fraassen, B. (1984) "Belief and the Will", *Journal of Philosophy*, pp. 235-56

Van Fraassen, B. (1995) "Belief and the Problem of Ulysses and the Sirens", *Philosophical Studies*, pp. 7-37

Weisberg, J. (2004) "Conditionalization, Reflection, and Self-Knowledge", manuscript