THE LESSER EVIL:
A DEFENSE OF SELF-SAMPLING AND CENTERED CONDITIONALIZATION

Daniel Kokotajlo

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Philosophy.

Chapel Hill
2017

Approved by:

John Roberts

Matt Kotzen

Carla Merino-Rajme

ABSTRACT

Daniel Kokotajlo: The Lesser Evil: A Defense of Self-Sampling and Centered Conditionalization
(Under the direction of John Roberts)

In this paper, I develop a taxonomy of views about how beliefs should be updated in situations of indexical uncertainty. Drawing on previous literature, I show how each view is subject to its own apparent *reductio ad absurdum*. I argue that two of the three views are more absurd than was previously believed. I argue that the third view—which is held by Bostrom, Elga, Sebens & Carroll, Arntzenius & Dorr, and myself—actually has a pretty good defense against the objection it faces. This matters because the various views have important empirical implications.

ACKNOWLEDGEMENTS

First, I would like to thank my adviser, John Roberts. He has met with me more or less regularly for more than a year now, and given me advice on all manner of topics, from how to phrase things in my thesis to how to achieve my life goals. I would also like to thank Carla Merino-Rajme and Matt Kotzen for serving on my committee, and for going above and beyond the call of duty in doing so. Their comments and suggestions guided this paper from its earliest stages. Finally, I thank my friends Aliosha Barranco Lopez, Alex Campbell, Sylvie Ramirez, Phil Bold, and most of all Ada Lin for making these last two years the best in my life so far.

TABLE OF CONTENTS

Introduction:

The Doomsday argument says that we should significantly increase our credence that humanity will go extinct in the next few hundred years.[1] The Fine-Tuning argument says that we should significantly increase our credence in Theism and/or in some sort of multiverse cosmology.[2]Anthropic Shadow arguments say that the probability of civilization-destroying events is higher than commonly believed—that tail risks of this sort are systematically more likely than trends in the data indicate.[3] Arguments from entropy say that most of our current cosmological theories are false.[4]

Formal epistemology constructs and compares formal models for how to reason, typically versions of Bayesianism.[5] Each of the arguments mentioned above has the interesting property of being sound on some of the models which have been thus far proposed, and unsound on others. Thus, the controversy over which version of Bayesianism (if any) is correct has important and wide-ranging implications.

[1] Because our observations (of our birth-rank) are more probable if there are fewer people in the world. (Leslie 1992)

[2] Because it would be incredibly unlikely for the constants of nature to be life-supporting by chance. (Tegmark 2014)

[3] Because extinction events and near-extinction events are underrepresented in our data. (Bostrom et al, 2010)

[4] Because they predict that we are probably freak observers, and hence are self-undermining. (Carroll 2017)

[5] For our purposes this means that they speak of credences, rather than beliefs, and paradigmatically say that you should proportion your credences according to some sort of conditionalization rule.

This paper classifies three popular Bayesian views about how to reason, explains how each kind of view faces its own purported *reductio ad absurdum,* and argues that a particular view—the Self-Sampling Assumption—can avoid the objection it faces. This paper is a response to a paper by Christopher Meacham, who invented some of these *reductios,* and who instead favors a model called Compartmentalized Conditionalization. (Meacham 1)

The structure of this paper is as follows. In the next section, which can be skipped by readers familiar with this literature, I define key terms and explain important concepts that will be used throughout. In Section 2, I lay out the three kinds of views—dubbed "Up-And-Down," "Down-Only," and "Hold-Steady"—and summarize the problems with them. This provides a roadmap and reference for the rest of the paper. In Section 3, I discuss the "Up-And-Down" view and the argument against it. In Section 4, I discuss the "Down-Only" view and the argument against it. In Section 5, I discuss the "Hold-Steady" view, the argument against it, Meacham's defense against that argument, and why the defense doesn't work. In Section 6, I return to the "Up-And-Down" view and examine the argument against it in more detail. I argue that there are two versions that need to be considered, one involving potential infinities and one involving actual infinities. In Section 7, I defend the "Up-And-Down" view against the potential infinities version of the argument. In Section 8, I defend it against the actual infinities version. Finally, in Section 9, I conclude with some general thoughts and takeaways.

I. Key Concepts

This section explains what worlds and centered worlds are, and the other basics of the Bayesian framework. It also explains how synchronic centered conditionalization works, which will be useful for understanding several of the views we will consider later. Readers familiar with these concepts are welcome to skip this section.

In this paper, "possible worlds," "centered worlds," and "worlds" are all *hypotheses* in the minds of the agents we are modelling.[6] A *centered world* is a maximally specific hypothesis; it takes a stance on everything, including on indexical questions like "what time is it?" and "who am I?" A *possible world* is maximally specific *except* that it doesn't take a stance on those sorts of questions. It only takes a stance on non-indexical matters; all its pronouncements are in the third person. Thus a possible world is a set of centered worlds—a set of centered worlds which differ only in their answers to questions like "who am I?" and "what time is it?" Different possible worlds contain different numbers of centered worlds; for example, a possible world in which the only people who ever exist are Adam, Eve, and God, and they only live for ten seconds each, contains thirty centered worlds.[7] Sometimes it is helpful to think of a possible world as a physical object rather than a hypothesis—just picture it as a universe, rather than as a description of one—and likewise it's helpful to think of a centered world as a particular person at a particular time in a particular universe.

In this paper I'll use "world" as shorthand to mean either possible world or centered world, depending on the context. In cases where it might be unclear, I'll use the full term. I'll use "center" as shorthand for "centered world."

Classical bayesian conditionalization decrees that your credences in any given proposition A should be P(A|E) where P(.) is the credence function you had a moment ago, and E is whatever

[6] That is, they are *epistemically* rather than metaphysically or logically possible.

[7] Assuming the relevant unit of time is the second. It's unclear what unit of time, if any, should be used in our models; fortunately, none of the claims in this paper depend on which unit we choose. There's an additional caveat: As I've set things up, it's conceptually possible for there to be more centered worlds than these in this world; for example, imagine a hypothesis which says "You are no one" instead of saying you are Adam, Eve, or God. Arguably such a hypothesis is coherent, though the versions of Bayesianism I am familiar with don't include such hypotheses.

new evidence you've gained since then. Synchronic Centered Conditionalization is basically the same thing, but with two tweaks:

> **Synchronic Centered Conditionalization (SCC):** For any given agent, at any given time, the rational credence to have in any proposition A is P(A|E), where E is the total evidence available to the agent at that time, and P(.) is the agent's *hypothetical prior credence function*. Propositions are sets of centered worlds.[8]

The first tweak is the "Centered" part. All that means is that we are explicitly dealing with credences in indexical claims as well as non-indexical claims; the hypothesis space is constituted by centered worlds. The second, more substantial tweak is the "Synchronic" part. Rather than making P(.) be whatever your credences were a moment ago and E be whatever new evidence you've gained since then, we say that E is *all* of your evidence, and P(.) is your *hypothetical prior credence function* — basically, a construct that means something like how inherently plausible you find the various hypotheses, before taking into account any of your evidence.

Synchronic Centered Conditionalization is appealing in its simplicity; it also has very nice properties such as expected-accuracy-maximization. (Das 2017) (Classic Bayesianism has some of these properties too, but leads to problems in e.g. Shangri-La cases.)[9] However, SCC is compatible with a wide range of actual behaviors, because it places no constraints on the hypothetical prior. Later in the paper we will consider versions of SCC that are built by adding in such constraints.

[8] I don't know who first came up with the idea. SCC sometimes goes by the name of (centered) Ur-Prior Conditionalization, sometimes (centered) Hypothetical Prior Conditionalization, and sometimes the Principle of Total Evidence. (Meacham 2008)

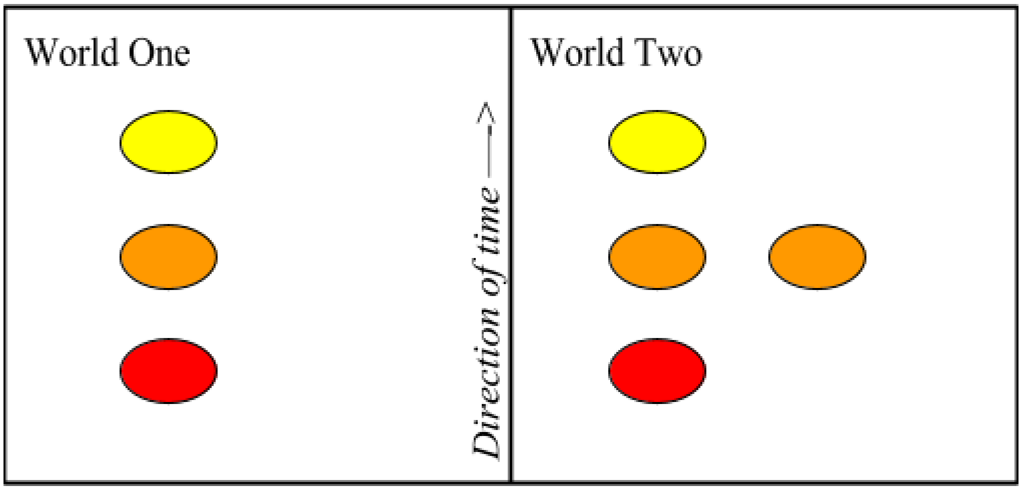
[9] See Arntzenius (2003)

II. The Options Available

This section maps out and explain the three categories of views that I'll be considering in this paper. I will also, for completeness' sake, discuss what it would take to have a view outside these categories.

Consider the following diagram, depicting two possible worlds that someone is dividing credence between: (It helps to think of it as a depiction of the Sleeping Beauty problem)[10]

Figure 1: Paradigmatic Case



Each disc is a centered world. Discs that share a color are in the same evidential state. This diagram does not specify whether there are additional discs in these worlds besides the ones depicted; however, we stipulate that there are no additional discs *of the same color* as the ones depicted here.

First the person is in the red evidential state, then the orange, and then the yellow. How should their credence in World Two relative to World One change over time? You might think it should go up and then down, or that it should stay the same and then go down, or that it should

[10] This diagram is an abstraction from the famous Sleeping Beauty case: World One is the heads world and Two the tails world; red is Sunday evening, orange is the evidential state of just having woken up during the experiment, and yellow is the evidential state of just having been informed that it is Monday. (I suppose there would also be a green circle in World Two, representing the state of just having been informed that it is Tuesday. I'm focusing on the red, orange, and yellow states though.)

hold steady throughout. The three types of views I consider in this paper are generalizations of these three options; each view is characterized by a thesis of the same name. This chart provides a helpful summary of the views and the problems with them, which I will describe later:

The Type of View	**Who Holds It?**	***Reductio* Argument Against It**
Up-And-Down Thesis: Increases in the relative number of centers in a world compatible with your evidence make you increase your relative credence in that world, and decreases make you decrease it.	Myself, Bostrom[11] (*Self-Sampling Assumption)*, Elga, Sebens & Carroll *(ESP-QM),*[12] Arntzenius & Dorr[13]	*Many-Brains Argument:* Your credence that the world contains many duplicates of you, and that you are not the original, will be constantly increasing.
Down-Only Thesis: Decreases in the relative number of centers in a world compatible with your evidence make you decrease your relative credence in that world, but increases don't change your credence at all.	Lewis, arguably many people who follow classical Bayesianism	*Sadistic Scientists Argument:* In the long run your credence that duplicates of you are being created and destroyed will approach 0, even if you are seeing it happen with your own eyes.
Hold-Steady Thesis: Neither increases nor decreases in the relative number of centers compatible with your evidence make a difference.	Meacham (*Compartmentalized Conditionalization*), Neal (*Full Non-Indexical Conditioning*)	*Varied Brains Argument:* Your credence that the world contains at least one center in each possible evidential state can never go down and will increase rapidly.

Important Note: The increases and decreases discussed in this chart are understood to be increases and decreases that don't involve zero. Everyone agrees that when the number of centers in your evidential state in a world is zero, you should rule out that world as incompatible with your evidence.

[11] Technically, Bostrom's view is "It Depends" (described below) because he uses centers-in-your-reference-class rather than centers, and because he leaves it wide open what your reference class is. However, there's certainly a *natural version* of Bostrom's view that fits in this category: The one in which your reference class is all centers.

[12] (Bostrom 2002)

[13] (Arntzenius & Dorr 2016)

What they disagree about is what to do when both of the worlds you are considering have at least one center which is in your evidential state, as depicted in Figure 1.

Other views, which will not be considered in this paper, are:

> **Up-Only:** Increases in the relative number of centers in a world compatible with your evidence cause you to increase your credence in that world, but decreases don't change your credence at all.

Nobody that I know of holds this view. More importantly, it is vulnerable to *both* the Varied Brains Argument and the Many Brains Argument.[14]

> **It Depends:** Sometimes increases in the relative number of centers in a world compatible with your evidence cause you to increase your credence in that world, but sometimes they don't. It depends.

This view is promising, but there are too many ways it could go: too many things "it" could depend on. So I won't consider it here, other than to make the following important remark: Each of the three arguments mentioned above—*Many Brains, Sadistic Scientists,* and *Varied Brains*—involves a peculiar skeptical hypothesis or situation designed to exploit the view. They can be tailored to fit the specifics of the view they are dealing with. As a result, it's not obvious that saying "It Depends" will help avoid the arguments, because the skeptical scenario can be modified to include whatever "it depends" on.[15]

> **Backwards:** Like one of the above views, except that increases in the relative number of centers in a world compatible with your evidence cause *decreases* in your credence in that world, etcetera.

I see no reason to hold this view; I mention it for completeness' sake.

> **Something Completely Different:** Perhaps all of these views are based on a mistake somehow. Perhaps we shouldn't have credences at all, for example, but rather beliefs.

[14] And the defense of the Both-Kinds view that I'll give does *not* work to defend the Up-Only view.

[15] For example, Bostrom's view makes it depend on whether the centers are in your reference class—so all we need to do is specify in the skeptical hypotheses that they are. There may be views that posit dependencies that allow them to truly wriggle free from these arguments. I can't really go into more detail here without explaining the arguments first, so I'll return to the subject at the end of the paper.

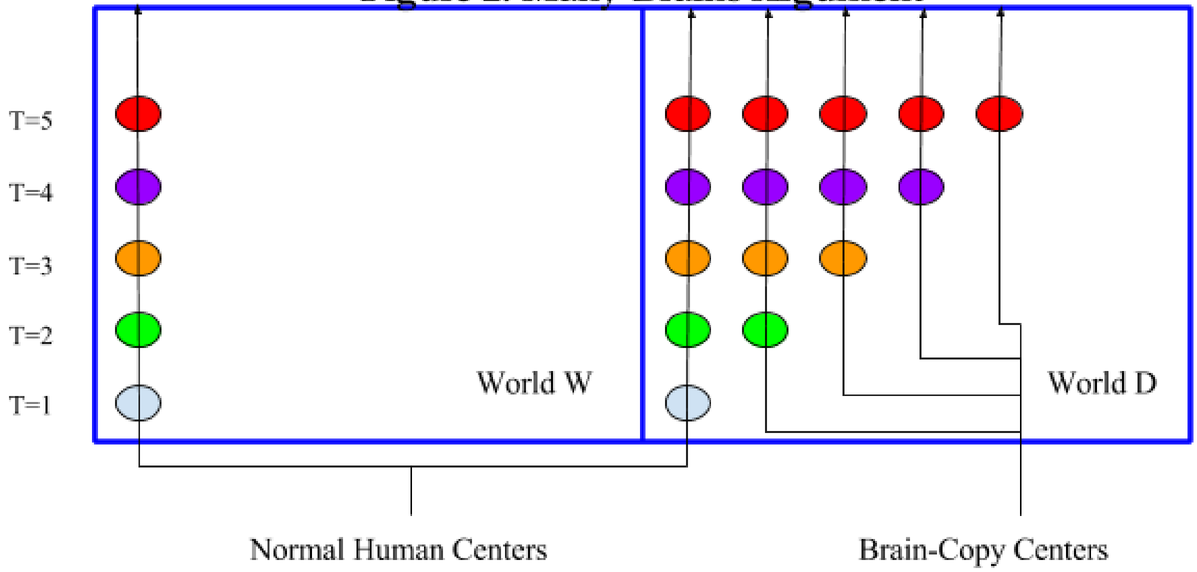
This view is beyond the scope of this paper. I mention it for completeness' sake.

In the next three sections, I will walk through the three rows of the chart, explaining the types of views in more detail and laying out the arguments against them.

III. Up-And-Down Views & the Many-Brains Argument

Meacham, who came up with the Many-Brains-Argument, targeted it at a much narrower range of views than this. (He introduced this argument as part of a larger attack on synchronic centered conditionalization[16]) However, it's easier to explain the argument at this level of generality:

Figure 2: Many-Brains Argument



Each dot is a centered world; the two blue boxes are possible worlds W and D. Dots that share a color are in the same evidential state. We are to imagine the pattern repeating for *t* > 5, though Meacham doesn't say for how long.

[16] Meacham's view—Compartmentalized Conditionalization—is, I suspect, equivalent to a particular version of synchronic centered conditionalization, namely the one you get when you add the following constraint on priors: The ratio of the total prior W1 assigns to evidential state E to the total prior W2 assigns to E, is the same for any E which is in both W1 and W2, where W1 and W2 are arbitrary worlds.

The argument goes like this. Pick your favorite non-skeptical hypothesis, whatever it is. Presumably it will look something like World W: A chain of centered worlds, representing yourself as you age.[17] Construct a skeptical hypothesis, World D, by taking your favorite hypothesis and then modifying it to include many identical copies of you being created far away, at a rate of one per unit of time, and maintained in such a way that they continue to be in exactly the same evidential state as you.

Since you subscribe to an Up-And-Down view, your credence in World D increases every time step, since the number of centers in D compatible with your evidence (relative to the number of centers in S compatible with your evidence) goes up. So even if you begin with almost zero credence in World D and almost complete confidence in S, given enough time, you will become arbitrarily confident in D relative to S.

Of course, in the actual world our hypothesis spaces are not so restricted; they contain more than just these two worlds. But it's plausible to think that this result generalizes: When we include all the hypotheses that we actually have credences in, we'll end up more and more confident over time in these strange skeptical duplication hypotheses. This seems bad.

I will hold off commenting on this argument for now, because sections 7, 8 and 9 will handle it with great detail. Section 6 will also explain why I interpret Bostrom's view as a Up-And-Down view and why, thus interpreted, it has initial plausibility. See Meacham for an account of why Elga's view fits the Up-And-Down category. (Meacham 14)

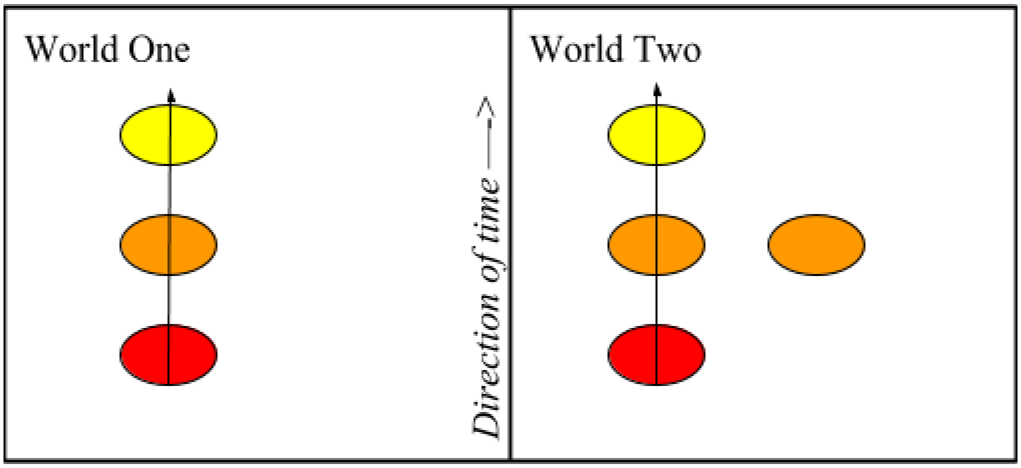
IV. Down-Only Views & the Sadistic Scientists Argument

This type of view holds that decreases in the relative number of centers in a world compatible with your evidence make you decrease your relative credence in that world, but increases

[17] There will be other chains corresponding to other people as they age, but I leave these out for simplicity.

don’t change your credence at all. Meacham has a good argument that Lewis holds this view. (Meacham 17) I also suspect that at least some forms of diachronic Bayesianism fit in this category. Here’s why:

Figure 3: Diachronic Reasoning?



Each disc is a centered world. Discs that share a color are in the same evidential state. Let’s stipulate that there are no additional discs in the worlds besides these ones. The arrows represent relations of personal identity across time. Suppose that nothing in particular is happening in either world: The subject is just watching a clock tick and thinking.

Technically, P(Orange|Red) is zero, so it’s hard to say what classical diachronic Bayesianism says you should believe when you are in the orange evidential state.[18] Yet when you were in the red state, you had complete confidence that you *would* end up in the orange state: Both World One and World Two predict that. So intuitively, when you find yourself in the orange state, you shouldn’t change your relative credence in World One or World Two from whatever it was before. Now, your credence in “I’m the center on the far right in the World Two diagram” should be nonzero when you are in the orange evidential state.[19] As a result, when you are in the orange evidential state you should think that the likelihood of seeing the yellow evidential state in a moment is higher if World

[18] The probability that “I am now in the orange evidential state” given “I am now in the red evidential state” is zero.

[19] After all, it is by definition compatible with your evidence. Surely it is unreasonable to assign *zero* credence to a hypothesis you are considering which is compatible with your evidence.

One is the case than if World Two is the case.[20] So when you are in the yellow evidential state, you should increase your credence in World One relative to World Two.

Thank you for indulging me in that digression. If it hasn't already paid for itself by helping to motivate the Down-Only view, I hope it will do so in this next part, by making it easier to explain the Sadistic Scientists Argument:[21]

Suppose you subscribe to the Down-Only view, and you have an enemy who wants to drive you insane and has the technology and resources to create duplicates of you. This enemy can ruin you by creating a duplicate of you, and then deleting it, and then creating another one, and then deleting it, and so on. (To visualize this, just imagine Figure 3 except with the pattern repeating for many more time steps into the future.) The finishing touch is for your enemy to give you lots of convincing evidence that this is what they are doing—perhaps they send you videos, signed letters from credible third-party witnesses, perhaps they even give you a tour of their brain-duplication machinery. (They should do this before, or soon after, they start the duplication process.)

Why would this ruin you? Well, notice how in the original case (described in Figure 3) you end up with a lower credence in World 2 than you started with. In this case, World 2 is the actual world; it's also the world that you have tons of evidence for. But given enough time, you'll come to disbelieve in World 2, no matter how high your initial credence in it was, because each round of duplication-and-then-deletion causes you to decrease your credence in World 2. So eventually you'll believe in alternative hypotheses in which this duplication is not taking place, despite all evidence to

[20] On some views of personal identity, this move can be blocked by saying that the two World Two orange centers are really one bi-located person, or else that for psychological continuity reasons the World Two yellow center is the successor state of *both* centers. This issue can be avoided by adding a green center to the diagram in World Two, psychologically continuous with the right-hand orange center.

[21] Meacham's original argument put it slightly differently: He merely said that we would become unreasonably confident that duplicates of us are *not* being created and destroyed, even if we were seeing it happen with our own eyes. For many people this is already an unacceptable consequence of the view.

the contrary. Perhaps you'll believe that you are hallucinating, or that the supposedly reputable third parties are part of a vast conspiracy against you, etcetera.

This seems bad. Sure, we don't have the technology to do this now, but what if we did? Anyone who subscribes to a view of this type would be vulnerable to an embarrassing sort of attack. Worse, there's reason to think that views of this type will eventually lose faith in most of our current best theories in cosmology, since most of them posit faraway duplicates being created and destroyed. (See Section 5 for further details.)

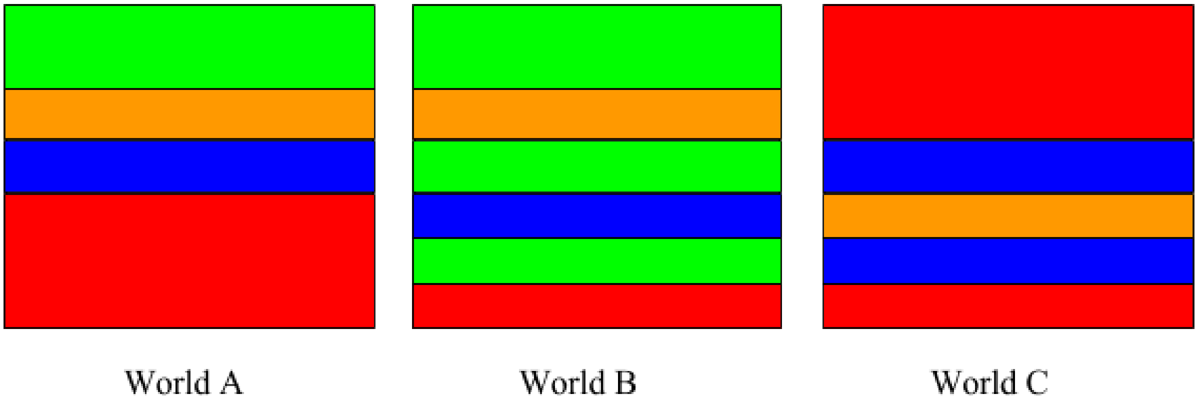
V. Hold-Steady Views & the Varied Brains Argument

In this section, I'll explain Meacham's view, and why it fits in this category. I will then give the Varied Brains Argument and explain why Meacham thinks it isn't so damaging. I'll finish by arguing that the argument is much more damaging than Meacham realizes.

Meacham's view is called *Compartmentalized Conditionalization*. (Meacham 5) It is a version of synchronic conditionalization; that is, it uses hypothetical priors and total evidence to tell each centered world what to believe without referencing the beliefs of earlier centered worlds. Each centered world, according to compartmentalized conditionalization, should choose its credences as follows: Rule out all the possible worlds inconsistent with your evidence, and renormalize the credences. Then, for each world that remains, rule out the *centered worlds* within it which are inconsistent with your evidence, and renormalize *within that world*. An example helps:

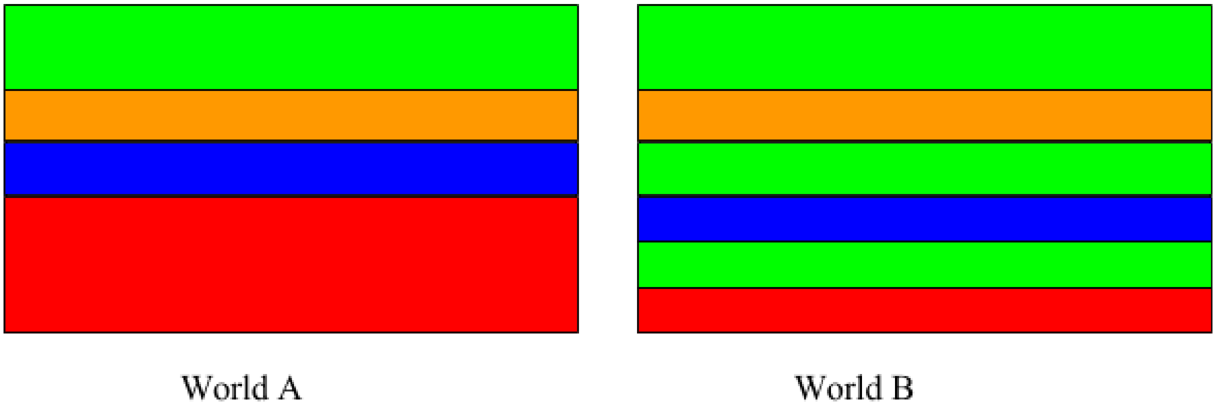
Suppose our hypothetical prior distributes probability according to the area of this diagram. Each rectangle is a center in a world; each color is an evidential state.

Figure 4: Hypothetical Prior



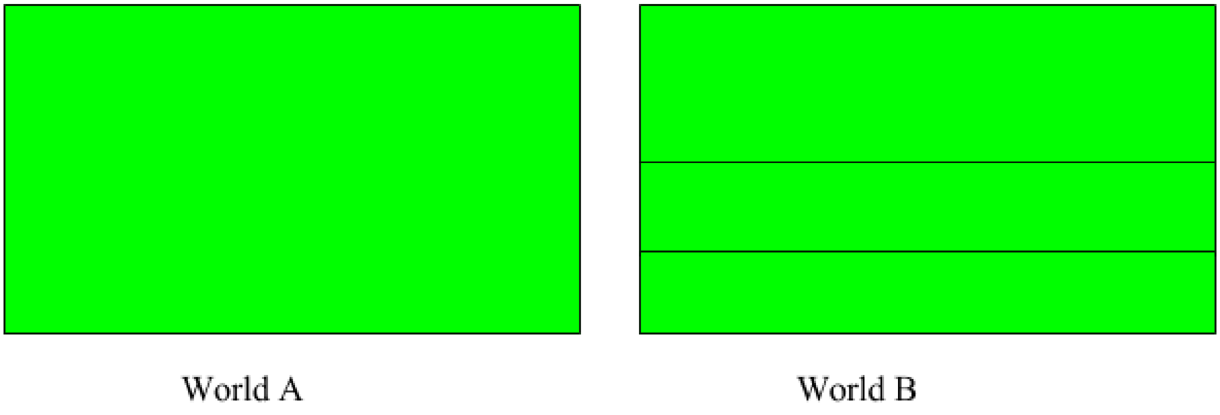
Suppose our current evidential state is Green. Then in the first step, we eliminate World C because it is incompatible with our evidence, and renormalize:

Figure 5: Compartmentalized Conditionalization, Step One



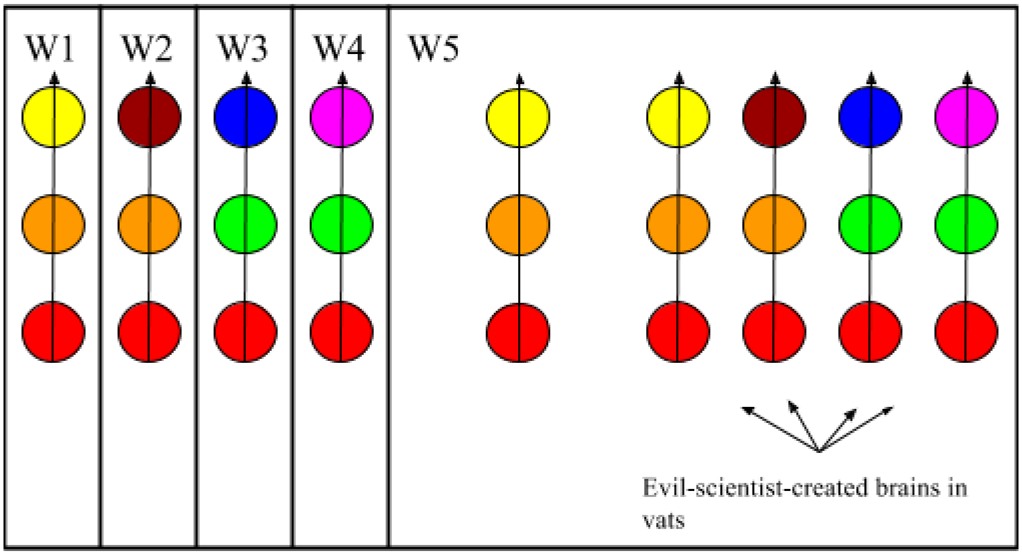
Next, within each world we eliminate the centers incompatible with our evidence, and renormalize within each world:

Figure 6: Compartmentalized Conditionalization, Final Step



Now that we understand how compartmentalized conditionalization works, we come to the Varied Brains Argument, paraphrased from Meacham: (Meacham 20)

Figure 7: Varied Brains



There isn't enough space in the diagram to show W6, W7, and W8, but they look like W5 except that their non-envatted centers mirror W2, W3, and W4 respectively.

Consider two sets of hypotheses. First, there is the set of normal hypotheses, which cover all the hypotheses about the world that you consider remotely plausible: W1-4. Then, there is the set of skeptical hypotheses, which are constructed from the first set by adding a mad scientist somewhere far away, who makes a bunch of duplicates of you *with a diverse array of experiences,* at least one for each series of centers in the first set.[22]

Anyone who follows compartmentalized conditionalization will, as time goes on, assign more and more credence to the second set of hypotheses relative to the first.[23] The rate of this growth will be quite dramatic; on plausible assumptions the credence in the second hypotheses grows by orders of magnitude every time step.[24] This is an unwelcome consequence.

Meacham takes this argument seriously, but he thinks it is not very damaging for his view. Here's what he has to say in defense of Compartmentalized Conditionalization:

> In the varied brains case, your credence in your strange worlds increases relative to your credence in your normal worlds because of the artificial way in which these doxastic worlds have been selected: all the strange worlds under consideration are ones that will end up matching whatever you experience, whereas many of your normal worlds won't match what you experience. If we restricted the normal worlds to those compatible with [what you in fact will experience], your credence in your strange worlds would not increase relative to your credence in your normal worlds. Likewise, if we placed no restrictions on which strange worlds were allowed, then [said experiences] would eliminate lots of strange worlds as well as lots of normal worlds. Whether your credence in strange worlds increases relative to your credence in normal worlds depends on which strange and normal worlds are your doxastic worlds which worlds our priors and evidence lead us to believe could be ours….
>
> 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. … The varied brains argument is a result of the second kind… The many brains

[22] Philosophers like to create thought experiments which begin like this: "Suppose a mad scientist builds a…" A common objection is that these scenarios are so unrealistic that we can't learn anything from them. I think that we philosophers should respond to this by changing all our examples to instead start with "Suppose a mad *philosopher* builds a… etc." This makes the thought experiment more realistic, because only a philosopher would ever want to do the things we allege mad scientists would do. (Why would a philosopher do those things? Because by actualizing these obscure possibilities, they would conclusively refute the common objection!)

[23] This is because none of the possible worlds in the second set will be ruled out, but some possible worlds in the first set will be ruled out, at every time step. So Step One of compartmentalized conditionalization will always be taking credence from the first set and distributing it to both sets. Step Two won't change this distribution, because it only redistributes credence *within* a possible world.

[24] Each time step you rule out all the normal worlds incompatible with your evidence; so, for example, if there are ten evidential states that you could experience in the next time step and they all have equal prior, then 90% of the normal worlds will be ruled out.

> 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.[25] (Meacham 21)

As I interpret him, Meacham's defense goes like this: Sure, in this toy model (where the hypothesis space contains only worlds in these two categories) we get the skeptical result. But we have no reason to think that this will generalize to our actual situation. After all, we consider many more hypotheses than these. When we include all the hypotheses we consider, there will be additional skeptical hypotheses, and perhaps many of them *will* be ruled out by our data, and so perhaps the total credence in skepticism will not rise over time—perhaps the two effects will cancel out.

A helpful analogy: On *any* Bayesian view, every time you observe something, call it A, you'll increase your credence in the skeptical hypothesis that an A-loving demon exists who is hell-bent on getting you to observe A. And if that's the only skeptical hypothesis you are considering, then your overall credence in skeptical hypotheses will rise too. But if you consider a wider range of skeptical hypotheses—as you should—then you needn't be worried, because the rise in credence in the aforementioned hypothesis will get will be cancelled out by the drop in credence in other hypotheses, such as the hypothesis that a B-loving demon exists who is hell-bent on getting you to observe B. Your overall credence in skepticism will hold steady.

As I interpret him, Meacham is hoping that something like this will apply in the case of the Varied Brains argument also.

In what remains of this section, I will argue that (an improved version of) the Varied Brains Argument is devastating for Compartmentalized Conditionalization and for Hold-Steady views more

[25] Meacham goes on to argue that the Sadistic Scientists Argument is also a skeptical problem of the second sort and thus also relatively benign.

generally.[26] There are several independent considerations, each troubling on its own, which together are quite damning.

I'll start with the fact that what Meacham calls "Strange worlds" are not actually that strange; the "people in outlandish situations" that Meacham pities are in fact *us*. This is because many of our current best cosmological theories are what Bostrom calls Big Worlds: They posit a very large universe, large enough that for any evidential state that you or I might experience, there is bound to be someone out there who is in that state. (Bostrom 51)

For example, if the universe lasts forever in a state of heat-death, random quantum fluctuations will, with very high probability, eventually produce observers in exactly our evidential state who are nevertheless massively deceived, their world being about to dissolve into chaos again. (Even ordinary non-quantum fluctuations have a high likelihood of doing this. Also, black holes are said to have some probability of spitting out any arbitrary object, so enough black holes lasting long enough would do the trick.) Another example: If some version of the Many-Worlds interpretation of quantum mechanics is true, then for every situation that you *might* end up in, there is some branch of the multiverse where you (or a duplicate of you, depending on how personal identity works) *do* end up in that situation. Another example: On some theories, the universe is spatially infinite; as Bostrom says, this would be the simplest topology. (Bostrom 51) Another example: If some sort of eternal inflation cosmological model is correct, then the universe is spatially and temporally infinite, and moreover contains infinitely large regions with every possible set of constants. If any of these theories are true, the world is Big: that is, the world is such that for every possible evidential state you or I might end up in, there is someone somewhere sometime who is in that state. (For a discussion of all of these ideas, see Tegmark, and also Bostrom)

[26] The improved version of the argument isn't original to me. Bostrom thought of it in 2002, though of course he didn't apply it to Meacham's view in particular. (Bostrom 2002) What I'm doing is elaborating on Bostrom's argument a bit and applying it to these particular views.

Now, you don't need to be convinced that we are *likely* in a Big World. As long as you assign, say, at least 0.000000001 credence to the proposition that the world is Big, you'll end up almost certain that it is, within a few seconds—if you follow Compartmentalized Conditionalization. This is because your evidence rules out large swathes of the non-Big worlds, but never rules out any of the Big Worlds. For example, in writing this paper I used a random-number generator to produce the following number between 1 and 100,000,000: 42,464,580.[27] Presumably your priors assign roughly equal credence to each of the possible numbers I could have displayed above.[28] So when you saw that number—42,464,580—you ruled out roughly 99.9999999% of the small-world hypotheses (weighted by their prior) and none of the Big-world hypotheses.[29] Obviously this sort of thing will make your credence in Big World approach 1 extremely rapidly, even if it starts out very small.

It gets worse. There's another problem for compartmentalized conditionalization: Your *relative* credences in various Big Worlds can never differ from the priors! This means that many ordinary parts of science become impossible. The following example illustrates this point:

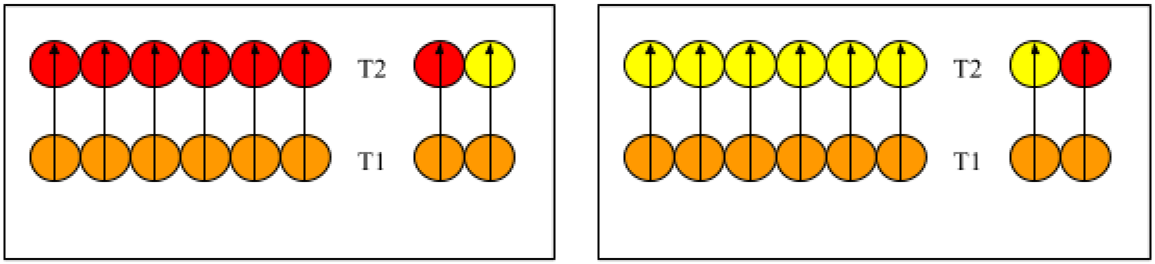
Cosmologists tell us that the cosmic microwave background radiation level (CMB) is about 2.7 degrees Kelvin. They discovered this by measuring it. Prior to measuring it, they were unsure about what it would be. But many of them thought we were likely in a Big World. So for them, the epistemic situation prior to measuring the CMB was like this:

[27] I used a "True random" number generator: https://www.random.org/

[28] Assuming you follow the Principal Principle or something like it.

[29] I'm ignoring for simplicity the medium-world hypotheses, in which the universe contains very large numbers of duplicates in a very wide variety of evidential states very similar to yours, but not *all* of the evidential states are represented. Compartmentalized Conditionalization would rule out many, but not 99.9999999%, of these worlds. Including these hypotheses would make the picture more complicated but would not fundamentally change the dynamics.

Figure 8: The Improved Varied Brains Argument



Big World 1: CMB = 2.7 K

Big World 2: CMB = 3.1 K

In this diagram, as usual, colors represent evidential states. Importantly, in this diagram orange represents the evidential state the scientists were in just prior to making the measurement of the cosmic microwave background radiation level, red represents the evidential state of just having observed a 2.7 K on the screen, and yellow represents the evidential state of just having observed a 3.1 K on the screen. Other possibilities besides these two worlds are left out for simplicity.

In the diagram, the twelve centers on the left-hand side of each world represent the vast numbers (proportions, if the worlds are infinite) of observers who genuinely are living on planets, who genuinely are observing the cosmic microwave background levels, etc. The four centers on the right represent the occasional freak observer, formed randomly out of heat-death or popped out of a black hole or something, whose randomly generated experiences and/or local environment happens to exactly match that of the normal observers, for now at least.

Consider how the scientists would handle this situation if they followed Compartmentalized Conditionalization. Suppose that at t=1 they have roughly equal credence in each of these two worlds, and that they are overall fairly confident that at least one of these worlds is actual. Then at t=2, they find themselves in the red evidential state. If they update using Compartmentalized Conditionalization, they end up with the *same* relative credences in these worlds. After all, none of these worlds were ruled out by the evidence! Even worse: They'd end up massively increasing their credence in the proposition "I am a freak observer," because their credence in the world where CMB = 3.1 is now entirely concentrated in the centered world in which they are freak observers.

In general, there are *two* problems for Compartmentalized Conditionalization. The first is what Meacham's Varied Brains argument was getting at, but more precise: Big Worlds are never ruled out by one's evidence, so a Compartmentalized Conditionalizer's credence that the actual world is Big can never decrease and in fact will increase rapidly.[30] If we take cosmologists at all seriously, the decrease in credence that non-Big skeptical hypotheses get will not be nearly enough to cancel out the increase that Big Worlds get.

The second problem is that a Compartmentalized Conditionalizer can never change their relative credences in the various Big Worlds. When they observe CMB = 2.7, for example, rather than decreasing their credence in Big World 2 and increasing their credence in Big World 1, they keep their credence in each world the same, and shift all of the credence *within* Big World 2 to the freak observers it contains—thus becoming much more confident that they are a freak observer! (A version of this problem arises in non-cosmological cases as well, if you don't like this talk about Big Worlds.)[31]

The two problems interact to make a pretty grim *reductio* of Compartmentalized Conditionalization: Anyone who holds it will become more and more confident that the world is Big each time they observe something that rules out a non-Big world, and moreover if we want to take contemporary cosmology seriously we should assign not-astronomically-tiny credence to the world

[30] More precisely: Since Big Worlds never get ruled out, credence in them will increase whenever non-Big worlds are ruled out. This will happen almost all the time--for example, worlds in which you have no duplicates get ruled out in droves every time step, since for every thing you might experience, there's a non-Big world which says you will experience that.

[31] Matt Kotzen articulated the example well: Suppose we duplicate you so that there are ten people in your evidential state. We flip a coin and on heads we say "heads" to nine people and "tails" to one, and on tails we say "tails" to nine people and "heads" to one. Compartmentalized Conditionalizers will continue to have 50/50 credence in heads/tails even after they are told, despite the fact that prior to being told they would have said "Conditional on heads, I'm probably going to be told heads, and conditional on tails, I'm probably going to be told tails."

being Big anyway, such that within a few *seconds* we should be almost certain that the world is Big. Moreover, once we become convinced that the world is probably Big, we are unable to do many routine scientific things like measure the CMB, and each time we learn about such measurements, we massively raise our credence that we are freak observers! (See appendix for further discussion.)

So far I've just talked about Compartmentalized Conditionalization, which is a specific view within the Hold-Steady category. Does the argument apply to every view of this type?

I think it does. The argument was originally designed to shoot down Full Non-Indexical Conditionalization (though not by that name, since the term hadn't been coined yet). FNC, promoted by Neal (2006) is the view that you should conditionalize on all your non-indexical evidence. (I suspect that it is equivalent to Compartmentalized Conditionalization.) So the only two views in this camp that I know of fall prey to this argument.[32]

Moreover, it seems like the reason why World 2 is intuitively disconfirmed by the red evidential state has something to do with the fact that the number of red evidential states in World 2 is so much smaller than the number of orange evidential states, and moreover the reason why World 1 comes out ahead is that, compared to World 2, World 1 lost very few evidential states in the transition from orange to red. Of course, these facts may not be the proximate justifications, but surely they are involved somehow. So a view of this type—a view which says that such facts are irrelevant—seems problematic.

[32] Nilanjan Das at one point proposed a "Relevance-Limiting Thesis" which is yet another example of a view in the neither-kind category that falls to this argument. The thesis states that changes in your evidential state that don't rule out two possible worlds don't change your relative credence in them. (Das, personal conversation)

VI. Defending the Up-And-Down View

Let’s step back a bit and assess the situation. I’ve discussed three prominent kinds of views about how to reason in response to evidence. I’ve explained one major problem for each of them, and I’ve argued that the problem for the Hold-Steady views is far worse than Meacham thought. In the remainder of this paper, I’ll argue that the problem for the Up-And-Down view is not as bad as Meacham makes it sound.

To do this, I’ll take up a particular view and show that it fits into the Up-And-Down category. I’ll then reexamine the Many-Brains Argument and split it into two versions, one involving potential infinities and another involving actual infinities. I will present a more or less “straight” solution to the first version in Section 7, and then in Section 8 I’ll say some things to dampen the force of the second version.

The view I’ll be defending is called SSA, for Self-Sampling Assumption.[33] It’s what you get when you combine Synchronic Centered Conditionalization with an indifference constraint on the hypothetical prior:

> **Synchronic Centered Conditionalization:** For any given agent, at any given time, the rational credence to have in any proposition A is P(A|E), where E is the total evidence available to the agent at that time, and P(.) is the agent’s hypothetical prior credence function.
>
> **Indifference Constraint:** Centered worlds that are part of the same possible world should get the same prior.

[33] I call it that because that’s what Bostrom calls it. I caution that my version isn’t exactly the same as Bostrom’s view, but it almost is. The major difference is that Bostrom says we should reason as if randomly selected *from our reference class* rather than from all centers in the world. He leaves it wide open what we should take our reference class to be. I don’t think this difference means much in this context, because I’ve left it wide open what we should take centers to be. A similar view is defended by Arntzenius and Dorr (2016).

SSA entails the Up-And-Down thesis.[34] It also an instance of the class of views that Meacham explicitly targets in his version of the Many-Brains Argument.[35] And it's quite a popular view: As best as I can tell, all of the versions of the Up-And-Down view that I mention in the chart in Section 2 are equivalent to or closely related to SSA. So if SSA can survive the Many-Brains Argument, then the Up-And-Down thesis has been successfully defended.

Recall the Many-Brains Argument: Since increases in the relative number of centers compatible with your evidence cause you to increase your relative credence, your relative credence in a skeptical hypothesis in which many duplicates of you are constantly being created will increase over time, eventually reaching unacceptably high levels. The argument is supposed to start with a toy case in which you only consider two possible worlds, and then generalize to our actual situation with many possibilities. This generalization is the step I wish to challenge.

In the next section I will show that, if your priors behave nicely, when we generalize to include all *finite* possibilities, the increases in credence that some skeptical hypotheses get will be exactly cancelled out by decreases in credence that other skeptical hypotheses will get, so that your overall credence in skepticism (and, correspondingly, in the normal, non-skeptical hypotheses) will stay the same. In Section 8, I'll explore what happens when we generalize further and include infinitely large possibilities. The situation here is much more murky.

[34] To see this, consider the original case (from Figure 1) and think about how SSA handles it.

[35] Meacham aims his argument at any view which follows SCC, Elga's Indifference Principle (which says that centers in the same evidential state & world get the same prior) and a Continuity Principle (which says that the ratio of prior in centers in the normal world to certain nonenvatted centers in the skeptical worlds should remain constant over time). (Meacham 24) The indifference constraint of SSA entails both of these two latter principles.

VII. Cancelling

In this section I'll talk about what happens when we generalize the Many-Brains argument to include every finite possibility—specifically, every possibility in which the number of centers is finite.

The Many-Brains Argument initially considers just one pair of hypotheses: A normal world and a skeptical world just like it but with increasing numbers of duplicates. When we generalize to include all the worlds, we'll add many more such pairs. But that won't be all that we add. In addition to including more normal worlds, and more skeptical worlds just like them but with increasing numbers of brains, we'll also include more skeptical worlds just like them but with *decreasing* numbers of brains, and also with *oscillating* numbers of brains, and so on. In fact, for every normal world, and for every function assigning numbers to evidential states in that world*,* there is a skeptical world just like that normal world except with numbers of duplicate brains determined by that function.

On an up-and-down view, your credence in skeptical hypotheses in which the number of duplicates is increasing (relative to your credence in the normal hypothesis from which it is derived) will be constantly going up. However, at the same time and for the same reason, your credence in skeptical hypotheses in which the number of duplicates is *decreasing* will constantly be going down. This might seem strange, but it really isn't: It's actually a standard feature of Bayesian views. (Recall the earlier example of the skeptical hypothesis in which there is an A-loving demon hell-bent on making you observe A. When you observe A, your credence in that hypothesis will go up, and your credence in the similar B-loving-demon hypothesis will go down. On any Bayesian view, skeptical hypotheses will be getting confirmed and disconfirmed all the time.) The important question is, will your credence in skeptical hypotheses *simpliciter* go up or not?

It can be proven that, for SSA at least, given a natural assumption about your priors, the answer is no:

> **Cancelling:** The skeptical worlds can be divided into equivalence classes, such that *if* your prior is distributed evenly across the sub-classes within each class, your credence in the skeptical worlds as a whole will not change at all over time (if you follow SSA). Moreover, if your prior is *not* distributed evenly, the difference in your posterior credence will be *at most* the difference in your prior. (Proof in appendix)

It's fair game to invoke a feature of our priors to defend against the *reductio,* because trivially every Bayesian view will yield absurd results for *some* priors. The important question is whether or not priors that we ought to have lead to absurd results.

So, thanks to Cancelling, if only our priors are (or should be) more or less evenly distributed within these equivalence classes of skeptical hypotheses, the Many Brains Argument just doesn't work (in the finite cases at least.) But *are* our priors like that? Should they be like that? Perhaps we should reject this antecedent.

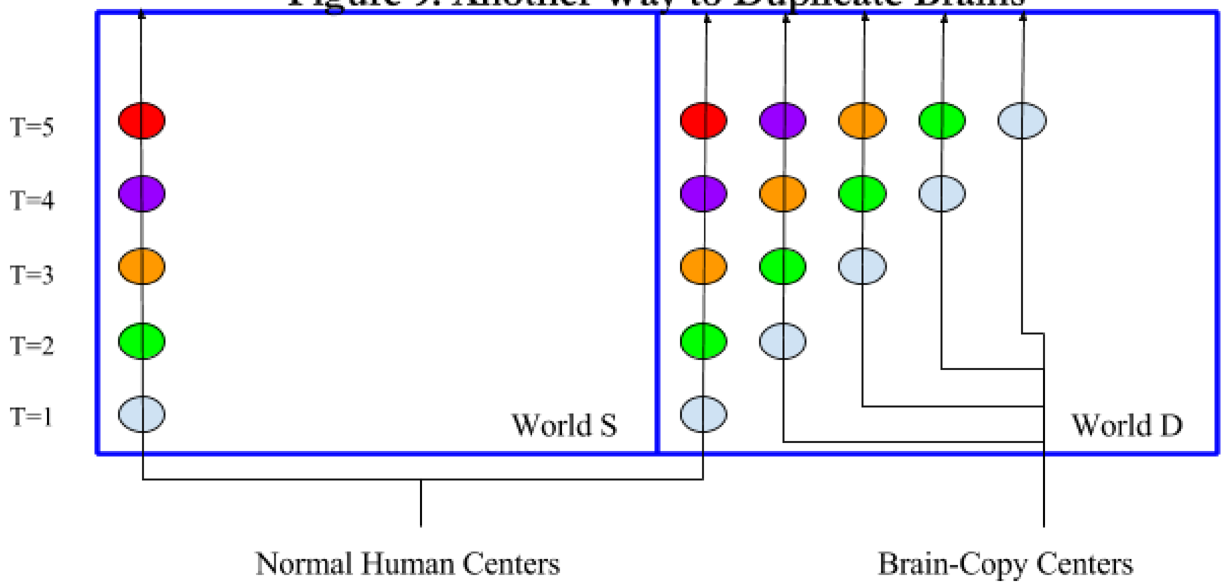
I don't have a positive argument in favor of the antecedent—I don't need one, since I'm on the defensive here. In the remainder of this section, I will defend it against an anticipated objection. But before I do, I wish to emphasize that it isn't enough to show that the antecedent is false; to make trouble for SSA we need to show that the antecedent is *systematically* false in a way that leads to absurd results for SSA. Continuing the analogy from before, suppose we successfully argue that a blue-loving demon really is more likely than a purple-loving demon because blue takes up more space in the range of light-wave frequencies or something. Have we thereby made a *reductio* of Bayesianism? No, because even though our credence in skeptical hypotheses as a whole might rise when we see blue, it won't rise by much, and if we see purple, it will shrink by a corresponding amount. The second part of Cancelling proves that similar reasoning will apply in the many-brains case.

One way to reject the antecedent would be to say that within the chain of centers that makes up a person over time, certain *times* are more likely to be duplicated (in greater numbers) than others. For example, you might think that later times are more likely to have more duplicates, perhaps because of some intuition about the direction of time and the way causal processes work. Suppose we find this much convincing: It's systematically easier to create more brains as time goes on than to create a bunch of brains at the beginning and slowly delete them.

This actually isn't enough to get the conclusion above, because centers at different times can have the same evidential state.[36] The issue is that there are two kinds of brain-duplication hypotheses that start small and grow larger: The first kind is the kind discussed so far, where at each time t, the duplicates are all in the same evidential state—so, when new brains are created, they are created in the same state as all the others. The second kind of brain-duplication hypothesis is one that creates new brains in the original evidential state, as depicted below:

[36] Earlier in the paper we dealt with simplified cases in which this was not true, but now we are relaxing that simplification.

Figure 9: Another Way to Duplicate Brains



Each dot is a centered world; the two blue boxes are possible worlds S and D. Dots that share a color are in the same evidential state.

These two kinds of worlds will cancel out, provided their priors are similar. I see no reason to think they aren't.

To conclude this section: While individual brain-duplication hypotheses may end up favored by your evidence, the group as a whole is not, unless your priors in the hypotheses in the group are systematically biased in certain ways. I've explored one way they could be systematically biased, and defused it. The overall conclusion is that, for finite hypotheses at least, the Many-Brains Argument just doesn't work.

VIII. Actual Infinities

In this section I'll talk about what happens when we extend the space of possibilities to include infinitely large worlds. (Specifically, worlds with infinitely many centers in them.)

Unfortunately, I can't invoke Cancelling here because my proof made use of the fact that the possibilities were finite. That doesn't mean it's false here—maybe the infinite possibilities will cancel out too—just that it hasn't been proven one way or another.

Part of the problem is that it is unclear what the right way to assign prior to centers in infinite worlds is. It's incoherent to divide it up evenly, as SSA and many other views say you should.[37] How to handle infinitely large worlds is a wide-open question and there's no reason to think that the eventual answer will vindicate the Many-Brains Argument.[38]

Besides, there *is* some reason to think that the eventual answer *won't* vindicate the Many-Brains Argument: There are probably multiple more or less reasonable ways to handle infinite cases.[39] As long as there is *at least one* reasonable way to handle infinite cases that *doesn't* vindicate the Many-Brains Argument, Up-And-Down views are safe, because they can choose to use that way.

All of this is, admittedly, quite speculative. So here's another, independent consideration: Notice that even within the infinite possibilities, our *lifespans* are finite in most of them. And, plausibly, *whatever* measure we choose over infinitely many centers, if there are only finitely many centers in our lifespan, the relevant sets of skeptical hypotheses involving different growth functions for those centers will cancel out.[40] So the only possibilities we need to worry about are those in

[37] Elga's Indifference Principle, alternatively known as the Weak Indifference Principle, is widely accepted and yet doesn't work in infinite cases. It says that your credence in two centered worlds which are in the same possible world and which are in the same evidential state should be equal. Even weaker indifference principles will still get the same result; consider an indifference principle which only applies to centers which are in exactly the same environments, with exactly the same causal histories, etc.--even this restricted principle will break down when applied to possible worlds that contain infinitely many parallel universes, all exactly the same, each one of which contains a center.

[38] By contrast, the Sadistic Scientists and Varied Brains arguments work more or less the same when we restrict ourselves to arbitrarily large finite hypotheses.

[39] That is, multiple ways to handle infinite cases that are about as reasonable as the alternatives. I expect that none of the ways will seem extremely compelling—that's precisely my point, in fact.

[40] It cancels out when there aren't infinitely many alien centers, so why would adding a bunch of other centers elsewhere in spacetime that are clearly distinguishable from me make a difference?

which we literally live forever, experiencing infinitely many different evidential states. Most of us assign very low credence to the proposition that we will live forever in this way. So the effect this will have on our actual beliefs will be extremely small—by stark contrast with Compartmentalized Conditionalization.

So I think that even if the Many-Brains Argument works in the infinite case—and it's not at all clear that it does—it won't make a noticeable difference in our credences until we are convinced (on the basis of ordinary evidence) that we have a significant chance of living forever, and even then, it will take a while. By contrast, the Varied Brains Argument applies to us already, given contemporary cosmological theories.

IX. Conclusion

There are more problems (challenges?) facing these views than the three *reductios* I've considered. And while I have defended the Up-And-Down thesis from the Many-Brains argument, my defense is not airtight; perhaps one day someone will successfully argue that we ought to have uneven priors in skeptical hypotheses, such that they wouldn't cancel out. Similarly, perhaps a more thorough exploration of the ways to handle infinities will bring back the Many-Brains argument as strong as ever. Nevertheless, I think I have shown that the Up-And-Down thesis is more promising than the Hold-Steady or the Down-Only theses.

The only alternative, as far as I can tell, is an *It-Depends* view: Perhaps we can do better than Up-And-Down views if we say that increases and decreases in credence are justified in some cases but not others, in a way that depends on more than just the relative numbers of centers-compatible-with-your-evidence. As I said before, the challenge for views of this sort is to specify the features-on-which-it-depends in such a way that none of the three arguments apply, without being completely ad hoc. This is difficult, because the arguments can be modified to add further features as needed. For example, you might say credence does sometimes go down when the relative number of centers compatible with your evidence goes down, but only in situations where experiencing something else was possible. This would avoid all of the *reductios* as they are currently stated, but the Sadistic Scientists argument could be trivially modified to apply.[41] I personally look forward to exploring this region of conceptual space; perhaps there really is a principled *It Depends* view which avoids all three arguments, and which avoids the Many-Brains argument in a more conclusive, airtight way than SSA does. We shall have to look and see. In the meantime, SSA (and the Up-And-Down thesis more generally) is the least bad option on the table.[42]

[41] To make the modification, just add that the duplicates created by the scientists are allowed to live a few seconds longer, but given different experiences than the original during those additional seconds. Now, at the time of a duplication, there are two possibilities for who you could be: the original and the duplicate, and because their experiences diverge in the future, there are two possibilities for what you are about to experience.

[42] Many thanks to John Roberts, Matt Kotzen, and Carla Merino-Rajme for helpful discussion and feedback on earlier drafts. Thanks also to Andrew Prudhon for conversation about the Cancelling proof.

X. Appendix

10.1 Proof that, for SSA, the expected value of the posterior in a duplicate world D_i equals the prior in D_i for any finite i, and likewise for the non-skeptical worldW_i. (Yes, this means it holds for *any* finite world. I don't know how to compute things for infinite worlds.)

The expected value of something is the weighted average of all the possible values of that thing, where the weight of each value is the prior probability of it obtaining. So the expected value of the posterior in Di is as follows:

$$EV(posterior(D_i)) = \sum_{t}^{i} \frac{weight(t) * P(D_i) * f_i(t)}{P(D_i) * f_i(t) + P(W_i)/i}$$

(Here we use f_i(t) instead of t/(pop of D_i), since we are considering the generalized case.) Meanwhile, the weight assigned to a given time/evidential state t is just the prior probability that that state would obtain, which is as follows:

$$\textit{weight(t)} = P(D_i) * f_i(t) + P(W_i)/i$$

Putting it together, a lot cancels out, and we get:

$$EV(posterior(D_i)) = \sum_{t}^{i} P(D_i) * f_i(t)$$

Now, by the definition of f_i(t), we know that:

$$\sum_{t}^{i} f_i(t) = 1$$

So the expected value of the posterior in D_i equals the prior in D_i.[43] Note that this is *not* the case for Compartmentalized Conditionalization; for a compartmentalized conditionalizer, the posterior in the proposition "The world is Big" will be higher than the prior *for every evidential state,* and thus the expected value will be too.

10.2 Cancelling, part I: Proof that the skeptical worlds can be divided into equivalence classes, such that *if* your prior is distributed evenly across the sub-classes within each class, your credence in the skeptical worlds as a whole will not change at all over time (if you follow SSA).

Let W_i be an arbitrary non-skeptical possible world. Define the $f_{ij}(.)$'s as the class of functions that distribute measure (summing to 1) across all the evidential states in W_i. Define D_{ij}as the class of all of the worlds which are like W_i but with additional duplicate centers, such that the prior that each world *d* in D_{ij}assigns to any given evidential state E is $P(d) * f_{ij}(E)$. (Unless we have reason to think duplicates of us are likely in the actual world, the class of allD_{ij}'s will consist entirely of skeptical hypotheses and hence have an extremely small total prior probability.)

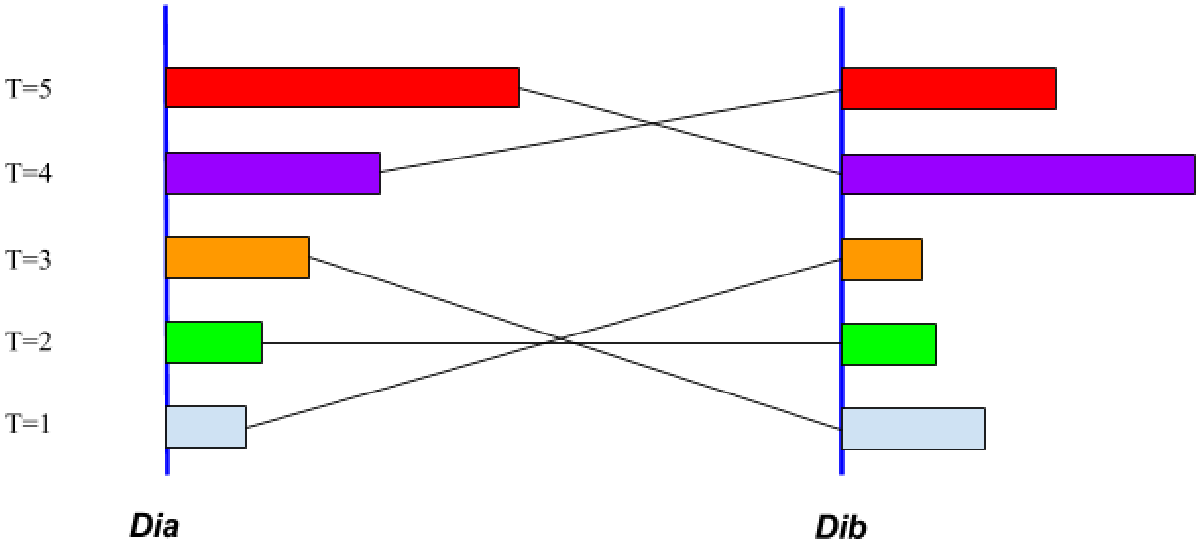
Definition: Two D_{ij}'s, $D_{ia}$and D_{ib}, are *permutation-equivalent* iff there is some mapping M(.) from the evidential states in $W_i$to the evidential states in W_i that is one-one and onto, such that

$$f_{ia}(t) = f_{ib}(M(t))$$

[43] The same proof can be easily modified to show this holds for Wi too.

Permutation-equivalence partitions the D_{ij}'s into equivalence classes.[44] Each equivalence class contains all and only the D_{ij}'s whose f_{ij}'s have the same structure. A diagram helps illustrate the idea:

Figure 10: Paradigmatic Example of Permutation-Equivalent Dij's



Colors represent evidential states; bars represent how much weight a given *fij()* assigns to that evidential state. Lines indicate the mapping relation M() that makes the two equivalent.

Now suppose that any two permutation-equivalent D_{ij}'s get the same prior probability. Then every equivalence class has a uniform prior, i.e. the prior of the equivalence class as a whole is evenly distributed among all the members of the class. (This is coherent because there are only finitely many members; there are only finitely many members because i is finite.)

It follows that each equivalence class, taken as a whole, will assign the same prior probability to each evidential state. (In other words, P(t1&C) = P(t2&C) for all t1, t2, where C is an arbitrary equivalence class of D_{ij}'s.)

[44] That is, the permutation-equivalence relation is reflexive, transitive, and symmetric.

Here is why:

The prior probability that an evidential state t gets from an equivalence class C is the sum, over all D_{ij}'s in C, of P(D_{ij})*$f_{ij}(t)$. The $f_{ij}(.)$'s in C consist of all the possible ways to reshuffle a particular series of bars, so to speak, over a list of colors. (Recall the diagram.) So the situation is perfectly symmetric for each color; the total length of all the bars they get will be the same. So each evidential state gets the same total prior, once you add up all the D_{ij}'s in C.

So, having established that each equivalence class assigns the same prior probability to each evidential state, we can conclude that D_i, the class of skeptical hypotheses as a whole, assigns the same prior probability to each evidential state: $P(D_i)/i$. (It is, after all, the aggregate of many such equivalence classes.) Now, W_i assigns the same prior probability to each evidential state too, $P(W_i)/i$, because there is only one center for each evidential state and SSA divides the prior evenly among all centers. So the relative posterior credence in W_i and $D_i$will be the same as the priors, for every evidential state.

10.3 Cancelling, part II: Proof that, if your prior is *not* distributed evenly, the difference in your posterior credence will be *at most* the difference in your prior.

Consider the worst-case equivalence class C, i.e. the one in which each *fij()* assigns 100% of its total probability mass to a single evidential state.[45] There are *i* evidential states, and thus we can divide this class into *i* sub-classes, each one choosing a different evidential state to concentrate on.

[45] Or arbitrarily close to 100%, if we want to stick with the earlier simplifying assumption that every evidential state in Wi also occurs in Di.

In this scenario, the prior that C assigns to a given evidential state E is simply the prior of the relevant sub-class of C that concentrates on E. So the likelihood of E given C is simply the prior of that sub-class of C that concentrates on E (divided by the prior of C).

So P(C|E) = P(E|C)*P(C)/P(E) = (prior of the sub-class of C that concentrates on E)/P(E)

Now, where W is any normal hypothesis, i.e. one without duplication,
P(W|E) = (1/i)*P(W)/P(E), by SSA. So the ratio of the posterior in C to the posterior in W is:
i*(prior of the sub-class of C that concentrates on E)/P(W)

Now, when the priors in C are evenly spread across each sub-class, the numerator will equal P(C), because the prior of each sub-class of C will be P(C)/i.
What happens when they are not? Well, suppose we multiply the prior in the sub-class that concentrates on E by a factor of M (and subtract prior from other sub-classes in C so that it balances out—notice that this means M can be at most i.)

Then the ratio of the posterior in C to the posterior in S (at E) will be multiplied by a factor of M too.

So in general, for this "worst-case" scenario, imbalances in the priors which prevent cancelling will cause exactly proportional imbalances in the posteriors. Notice that even in this worst-case scenario, the imbalance in the posteriors only happens at one particular evidential state; when you average over all evidential states there is no imbalance.

Now, I've been calling this the worst-case scenario without justification so far. But consider: No equivalence class is more extreme than this one when it comes to imbalances in the priors leading to imbalances in the posteriors; any other equivalence class will contain $f_{ij}(.)'s$ that assign probability to more than one evidential state, for example, and so the posterior credence in any given evidential state will be more than just the credence assigned to it by one sub-class, and so multiplying the prior of one sub-class by M will affect the posterior by less than M.

10.4 Discussion of Meacham's Defense of Compartmentalized Conditionalization:

Recall that, as I interpret him, Meacham's defense against the Varied Brains argument is somewhat similar to my Cancelling-invoking defense against the Many Brains argument. He seems to be hoping that when we add in the additional skeptical worlds, there *will* be skeptical worlds ruled out by our ordinary experiences, and they'll be ruled out in sufficient numbers to balance out the fact that the Big Worlds are not ruled out. In this appendix section I will summarize the reasons to think that this strategy doesn't work very well for Compartmentalized Conditionalization. I'll go into somewhat more detail than I did in the main text.

First, Cancelling doesn't apply, because the proof involves a certain notion of permutation-equivalence that just doesn't work in the case of Compartmentalized Conditionalization. The alternative worlds Meacham will need to invoke are going to be worlds that have *no* centers in various evidential states, not worlds that permute the numbers of centers in a given list of evidential states. So at best an analogue of Cancelling might apply.

Second, as I showed in 10.1, for SSA the expected value of the posterior in any hypothesis or set of hypotheses equals the prior, but for Compartmentalized Conditionalization, the posterior in any Big World hypothesis is greater than the prior, and so the expected value is too. That's weird.

Third, even if something like Cancelling works for Meacham (or doesn't work for me) such that the two views are on a par in this respect, there's still the other problem for Compartmentalized Conditionalization, namely the problem about being unable to change your relative credences in the different Big Worlds.

Fourth, and most importantly, for the Cancelling strategy to work, there must be a big stock of skeptical hypotheses which get billions of times more credence than the Big Worlds, which can be sacrificed to balance out the gain that the Big Worlds get. So for the Cancelling strategy to work, you have to be able to look cosmologists in the eye and tell them that not only are their beliefs wrong, they are so wrong that a certain class of skeptical hypotheses is way more plausible!

REFERENCES

Arntzenius, F. (2003) "Some Problems for Conditionalization and Reflection." *The Journal of Philosophy,* Vol 100, no 7, pp. 356-370

Arntzenius, F. and Dorr, C. (2016) Self-Locating Priors and Cosmological Measures. [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/11864 (accessed 2017-03-29).

Carroll, S. (2017) "Why Boltzmann Brains are Bad." Available online at https://arxiv.org/abs/1702.00850.

Bostrom, N. (2002) *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, 270 Mason Ave, New York NY 10016.

Bostrom, N, Cirkovic, M. and Sandberg, A. (2010) "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks." *Risk Analysis,* Vol. 30, No 10.

Das, N. (2017) "Justifying Ur-Prior Conditionalization." Unpublished manuscript.

Leslie, J. (1992) "Doomsday Revisited." *Philosophical Quarterly* 40(158):85-87.

Meacham, C. (2008) "Sleeping Beauty and the Dynamics of *De Se* Beliefs." Philosophical Studies, 138: 245-269

Neal, R. (2006) "Puzzles of Anthropic Reasoning Resolved using Full Non-indexical Conditioning." Technical report No. 0607, Department of Statistics, University of Toronto.

Tegmark, M. (2014) *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Vintage Books.