Existential Bias

Casper Storm Hansen

Abstract: To ascertain the rational credences for the epistemic agents in the famous cases of self-locating belief, one should model the processes by which those agents acquire their evidence. This approach, taken by Darren Bradley (*Phil. Review 121*, 149–177) and Joseph Halpern (*Ergo 2*, 195–206), is immensely reasonable. Nevertheless, the work of those authors makes it seem as if this approach must lead to such conclusions as the doomsday argument being correct, and that Sleeping Beauty should be a halfer. I argue that this is due to an implicit existential bias: it is assumed that the *first* step in those processes is the determination that the agent in question must necessarily exist. It is much more reasonable to model that determination as contingent and a result of other, earlier, steps in the process. This paper offers such alternative models. They imply an endorsement of what has mockingly been called "presumptuous" reasoning, and a massive shift of credences in favor of (1) the existence of a multiverse and (2) the Everettian interpretation of quantum mechanics.

This paper is concerned with four problems of self-locating belief: the doomsday argument, Sleeping Beauty, one that concerns the stochastic versus Everettian interpretation of quantum mechanics, and the fine-tuning argument. Bradley (2012) and Halpern (2015) have approached all these problems and two of them, respectively, by in effect asking how best to model the procedure that led to those scenarios' agents having the evidence that they have. I agree that this is the right question to ask. However, I also think that Bradley's and Halpern's attempts at answering it leave a lot to be desired. By identifying an existential bias in their reasoning, and then explaining how to avoid it, I will try to do better.

In the literature, each of these problems has received considerable attention *individually*, but there have been hardly any attempts to treat them simultaneously: in addition to Bradley and Halpern, I am only aware of Bostrom's (2002) and Friederich's (2021). That is unfortunate, because for each problem individually, it is not difficult to come up with some premises that support whatever conclusion fits one's preconceived intuition. However, the same premises are likely to either give a counter-intuitive result for one or more other problems, or appear to be ad hoc. Thus, the *real* problem is to come up with an overall theory that one can defend in its entirety.

Bradley and Halpern understand that, but I will argue that their theory cannot be defended, not because of its counter-intuitive results—mine has some as well—but for more fundamental reasons.

In a nutshell, the contentious issue is: what evidence can legitimately be employed for Bayesian conditionalization? Unlike Bradley, Halpern, Bostrom, and Friederich, my answer, in another nutshell, is: all of it!

1 The four problems

This section introduces the four problems. Halpern's treatment of the first two are discussed in the next section, and Bradley's of all four in section 3. My own account of them is then provided in section 4.

Doomsday: The doomsday argument is as follows. If the human race will continue to exist far into the future, then it will probably spread to other planets and solar systems, and the number of people will multiply far beyond the current tally. Hence, my birth rank, i.e., my place in the sequence of all humans who ever live ordered by time of birth, will be conspicuously small relatively to the average. Because it is unlikely that I am so special, it is unlikely that the human race will continue to exist far into the future. Or, at least, I should consider such continued existence less likely, after considering this argument, than I did previously (Carter 1983; Leslie 1996).

Without losing anything essential, we can streamline the discussion of this argument by assuming that there are only two options: either there will be one trillion people in the history of the universe, or two trillion. And I know this. Given this simplification, the question is how it should affect my credences for the propositions that there will be one versus two trillion people in total, if I learn that I am among the first trillion people.

Sleeping Beauty: Sleeping Beauty learns on Sunday that over the following two days, she will either be woken up once (on Monday), or twice (once on Monday and once on Tuesday). She will not be able to distinguish between a Monday awakening and a Tuesday awakening, in part because she will be given a drug after the Monday awakening that makes her forget it. The number of awakenings is determined by a fair coin toss: one if Heads, two if Tails. When she is woken up, what credence should she assign to Heads? *Halfers* say $\frac{1}{2}$, while *thirders* say $\frac{1}{3}$. A simple argument for the former position is that Beauty has not learned anything relevantly new when she wakes up, and therefore must have the same credence as on Sunday, at which time her credence should obviously be $\frac{1}{2}$; while an equally simple reason for believing the latter is the fact that if the experiment is repeated many times, only a third of all awakenings will be associated with Heads (Elga 2000).

Quantum Mechanics: According to the stochastic version of quantum mechanics, when a measurement is made on a particle in a superposition of Up

and Down spin, there may be a 50% chance of an Up outcome and a 50% chance of a Down outcome. Everettian quantum mechanics, on the other hand, holds that both outcomes will happen, but in separate branches into which the world splits. While these two accounts make no difference with respect to my subjective experience, they disagree about objective probabilities: the existence of an Up world has probability $\frac{1}{2}$ under the stochastic version, and 1 under the Everettian version. Does that mean that when I observe Up, the Everettian version is confirmed (Page 1999; Bradley 2011)?

To make this case as specific and simple as possible, it will be assumed that both versions of quantum mechanics have prior credence $\frac{1}{2}$, and that it is Up rather than Down that I observe.

Fine-Tuning: Our universe is fine-tuned for the existence of life: there are many physical constants that, if they had been slightly different from what they are, would have made life impossible. The existence of life thus seems like it was unlikely. However, for all we know, there could be multiple universes, whose physical constants vary. Thus, our existence might be due to the existence of many universes of which one happened to have the right constants, rather than extreme luck with the properties of a unique universe. So, should I regard my existence as confirmation that there are multiple universes (Leslie 1989)?

I make two simplifying assumptions for this case. First, the only possibilities are that there is one universe, and that there are two universes. Second, if a universe's physical constants are such that it is hospitable to life, then it actually contains life.

2 Halpern

I will first comment on Halpern's treatment of Doomsday, and then on his treatment of Sleeping Beauty.

Doomsday: Halpern points out that how I should update my credences on the information that I am among the first trillion people should depend on which protocol I should assume that "Nature" has employed to "decide" both how many people there will be in the history of the universe, and which one of them *I* am. He considers two such protocols.[1]

According to protocol 1, Nature *first* made the decision about whether there should be one or two trillion people, and it did so via a stochastic process in which each option had probability $\frac{1}{2}$. *Then* it decided who I should be, and it did so in a stochastic process in which the probability was uniformly distributed among all the available birth ranks: i.e., either among one trillion

[1] I am making some inconsequential changes to the presentation in the interest of simplicity and uniformity.

birth ranks or among two trillion, depending on the first decision. If Nature uses protocol 1, then, the probability that I am among the first trillion people on condition of there only being one trillion people in the history of the universe is 1, while the probability that I am among the first trillion people on condition of there being two trillion people in the history of the universe is $\frac{1}{2}$. It follows that the prior probability for the universe ending after one trillion people is $\frac{1}{2}$, and the posterior, after I have learned that I'm among the first trillion, is $\frac{2}{3}$.

According to protocol 2, on the other hand, Nature made the decisions in the opposite order. First, it decided on my birth rank via a stochastic process in which the probability was uniformly distributed among *two* trillion possibilities. Then, if my rank was determined to be less than or equal to one trillion, it decided whether there should be one or two trillion people, via a stochastic process in which each option had probability $\frac{1}{2}$. However, if my birth rank was larger than one trillion, Nature's hands were tied: there had to be two trillion people. Under this protocol, then, the posterior probability for the universe ending after one trillion people is $\frac{1}{2}$, while the prior was just $\frac{1}{4}$. Thus, on both protocols, my being among the first trillion indeed confirms that doomsday will arrive early.

There is an obvious problem with both of these protocols as analyses of the scenario: they both guarantee my existence. In both cases, it is as if Nature's *very first* decision was that I, privileged among all possible individuals, *must* come into existence, and that the rest of the protocol must therefore ensure that outcome. I would assume the opposite: i.e., that my existence is contingent. Hence, when we analyze Doomsday in terms of protocols that guarantee my existence, we run the risk that our conclusions are due to what might be called an existential bias. Just as confirmation bias consists in failing to adequately ascertain the evidential relevance of some of one's knowledge—namely, that which goes against one's beliefs—existential bias consists in failing to adequately ascertain the evidential relevance of one's knowledge of one's own existence.

Halpern thinks that protocol 2 is the most appropriate as a model, and argues for that position as follows. "[T]he primary choice is who you are. Once you exist, how long the universe will survive is only one of many questions that you could have asked" (page 200). This, I would counter, is based on an equivocation. My existence is a prerequisite for me being able to ask questions, and also for me being able to assign probabilities and conditionalize on evidence. However, it being a prerequisite for me considering evidence does not imply that it is a prerequisite for Nature making other choices; nor does it imply that my existence is not, itself, evidence that should be taken into account.

Below, I will propose a third protocol, free of existential bias, as a more

reasonable model.

Sleeping Beauty: Halpern also considers two different protocols for the case of Sleeping Beauty. According to protocol 1, Nature first decides the outcome of the coin toss, and then decides when *now* is. The first decision is made via a stochastic process in which both Tails and Heads have probability $\frac{1}{2}$. If its outcome is Tails, a second stochastic process decides between Monday and Tuesday, splitting the probability of $\frac{1}{2}$ for Tails into $\frac{1}{4}$ for Tails&Monday and $\frac{1}{4}$ for Tails&Tuesday. On the other hand, if the first outcome is Heads, Nature must let *now* be Monday.

According to protocol 2, Nature first decides whether *now* is Monday or Tuesday, with equal probability. Then, it decides whether the coin comes up Tails or Heads. According to Halpern, this results in each of the possibilities Tails&Monday, Tails&Tuesday, and Heads&Monday having probability $\frac{1}{3}$. While he is not explicit about it, I take it that the protocol is supposed to deliver that result through the following three steps. First, Monday and Tuesday are each assigned probability $\frac{1}{2}$. Second, those probabilities are each split in half, resulting in each combination of a day and a coin toss having probability $\frac{1}{4}$. And third, conditionalizing on the fact that she is awake, Sleeping Beauty ends up assigning probability $\frac{1}{3}$ to each of the three options that are consistent with that.

This two-protocol analysis is problematic, because there clearly aren't two different ways, corresponding to the two protocols, that the Sleeping Beauty scenario could be run. It is a single, clearly defined scenario. Thus, the strategy of trying to resolve the dispute between halfers and thirders by analyzing the scenario in terms of protocols would seem to be ineffective: it doesn't definitively come down on one side or the other. It would have to follow from the usual kind of description of the scenario (like the one I gave above) that Nature decides what day it is *before* it decides on the outcome of the coin toss, or the other way around. And it doesn't seem to follow.[2] Part of the problem is that how "before" should be interpreted is obscure. It is clearly not in a straightforward temporal sense, given that one of the decisions is about when *now* is. Also, one of the few things there is universal agreement about in the Sleeping Beauty debate is that it doesn't matter whether the coin toss happens Sunday, Monday, or Tuesday morning.

But the problem of the protocol approach failing to produce a clear verdict is even worse than it initially appears. This is because there are not just two contenders for the title of correct protocol analysis, but at least four. This is easiest to see if we use diagrams. Protocols 1 and 2 can be illustrated as in the top half of Figure 1. From these diagrams, it can be seen that the

[2] All Halpern has to say on the matter is the following (203-4): "To me, the first protocol seems more reasonable—it seems more consistent with the presentation of the story to think of the coin as being tossed first. But again, reasonable people can disagree."

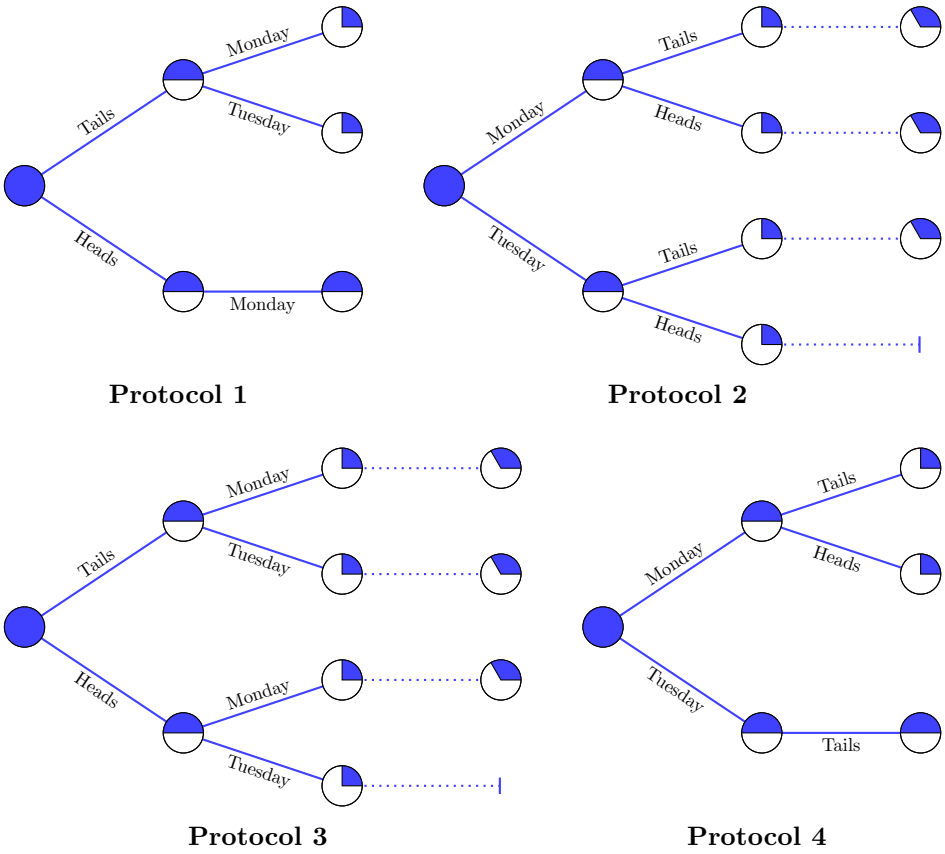

Figure 1 – Four protocols for Sleeping Beauty. Solid lines are used for Nature's assignment of probabilities, while dotted lines indicate conditionalization on Beauty being awake.

two analyses of Sleeping Beauty differ in *two* ways. First, there is the issue Halpern focuses on: whether Nature's first decision (in some sense of "first") is about (1) the coin toss or about (2) the day. Second, there is the issue of whether combinations of facts that are ruled out by Beauty's evidence should (a) not be assigned any probability from the outset, or should (b) receive probability in the form of hypothetical priors[3] that is then "retracted" by conditionalization. Halpern combines (1) with (a) in protocol 1 and (2) with (b) in protocol 2. But why not combine (1) with (b) in a third protocol, and (2) with (a) in a fourth protocol, as shown in the bottom half of Figure 1?[4]

It is not my intention to advocate for either of these combinations. (Indeed,

[3]The concept of hypothetical priors was first considered (but rejected) by Glymour (1980) as a way to handle old evidence, and subsequently endorsed by Howson (1991).

[4]Similarly, four protocols can be formulated for Doomsday. In that case, however, Halpern does not consider both of the "dimensions" along which the protocols can be varied.

the (2)/(a) combination results in an obviously unreasonable "quarter" position.) Rather, my point is that, in addition to the protocol-analysis strategy being insufficient by itself to deliver a verdict (which Halpern is aware of), that strategy plus a determination about whether Nature decides on the coin toss first or the day first is also insufficient.

At least three of the four protocols must be inadequate as analyses of the scenario. To make any progress, then, we need a method with which to rule out some such analyses. Below, I will argue that in fact all four are inadequate, and propose a fifth.

3 Bradley

Bradley (2012) has taken what looks, on the surface, like a different approach to the self-locating problems. He discusses them in terms of *selection effects*, and warns us of the dangers of not taking such effects properly into account when calculating probabilities. He opens with Eddington's (1939) classical cautionary tale of what happens if one naively draws inferences about the population of fish in a lake from a sample containing only large specimens, when that sample was caught using a wide-meshed net. This overall approach is one I approve of, but—as in the case of Halpern—I am critical of the details, for Bradley's conclusions also suffer from an existential bias.

Bradley compares the four problems of self-locating belief to some very simple scenarios in which a sample of one ball is drawn from an urn. In all the scenarios, the urn either contains one or two balls; each ball is small or large; and the contents of the urn are determined by at least one fair coin toss. In some cases, Heads and Tails in this first coin toss each imply one particular population of balls; in other cases, further coin tosses are needed. (But below, "Heads" and "Tails" only refer to the outcome of the first coin toss.)

The scenarios also differ along another dimension: the selection procedure by which a ball is drawn from the urn, after it has been determined what it contains. In some scenarios, the sample is drawn using a random procedure, i.e., one in which each ball in the urn has the same probability of being drawn. In others, a (maximally) biased procedure is employed: if the urn contains at least one small ball, a small ball will be extracted as the sample.

There are four combinations of scenario properties that are relevant. For each of them Bradley asks the question, "if a small ball is drawn, does that confirm Heads, confirm Tails, or confirm neither?" The four combinations and corresponding answers to that question are as follows:

1. If

- Heads implies that the urn contains one small ball,
- Tails implies that the urn contains one small ball and one large ball, and
- a small ball is drawn using the random procedure,

then Heads is confirmed. This is because the resulting sample is certain given Heads, but only has probability $\frac{1}{2}$ given Tails.

2. If

   - Heads implies that the urn contains one small ball and one large ball,
   - Tails implies that the urn contains two small balls, and
   - a small ball is drawn using the biased procedure,

   then nothing is confirmed. This is because the resulting sample is certain either way.

3. If

   - Heads implies that the urn contains one small or one large ball with equal probability,
   - Tails implies that the urn contains one small ball and one large ball, and
   - a small ball is drawn using the random procedure,

   then nothing is confirmed. This is because the resulting sample has probability $\frac{1}{2}$ either way.

4. If

   - Heads implies that the urn contains one small or one large ball with equal probability,
   - Tails implies that the urn contains two balls, each being small or large with equal and independent probability,
   - and a small ball is drawn using the biased procedure,

   then Tails is confirmed. This is because the resulting sample has probability $\frac{3}{4}$ given Tails, but only $\frac{1}{2}$ given Heads.

In that order, the four urn scenarios correspond to Doomsday, Sleeping Beauty, Quantum Mechanics, and Fine-Tuning, according to Bradley. I will discuss them in the same order.

Doomsday: In the case of Doomsday, a small ball represents the first trillion people, and a large ball the second trillion. The ball that is drawn represents

the group of which I am a part. Heads corresponds to there being one trillion people in the history of the universe, and Tails to two trillion. Thus, Doomsday corresponds to an urn scenario in which Tails implies that the urn contains one small and one large ball, and—since there cannot be a second trillion without a first trillion—Heads implies that the urn contains a small ball. And because I'm not guaranteed to be among the first trillion, but could (if there will be two trillion people in total) equally well be among the second trillion, Doomsday also corresponds to the case where the sample ball is drawn using a random procedure. Hence, according to Bradley, Doomsday is like the first urn scenario, so if I learn that I am among the first trillion people, that confirms that there will only be one trillion people: the prior probability was $\frac{1}{2}$, and the posterior is $\frac{2}{3}$.

While there are superficial differences between Halpern's and Bradley's frameworks of analysis, it is not difficult to see that Bradley's (implicit) assumptions can be formulated in Halpern's terms: namely, as assumptions about the order in which "Nature" makes the decisions that determine how the world is, who I am, and what evidence is available to me. In the case at hand, Bradley implicitly assumes that Nature made its decision about the total number of people before it decided who I am, for he is assuming that the latter decision depends on the former. That is, according to him, my being among the first and the second trillion people are both genuine possibilities if and only if Nature first decided on a total of two trillion people, whereas if there are only going to be one trillion in total, it is forced to make me one of that first/only trillion. But why is that? Why couldn't Nature, having first decided on a total of one trillion people, then decide with probability $\frac{1}{2}$ that I should be among them, and with probability $\frac{1}{2}$ that I should not exist? Apparently, because Nature's *very first* decision was that I should exist (as illustrated in Figure 2). Because I believe this to be an unreasonable assumption, I judge Bradley's conclusion to be affected by existential bias.

Bradley's conclusion rests on a deceptive analogy. In the world of urns and balls, we *could*—even though it seems silly, and clashes with a central convention of a long tradition of urn examples—first pick either a small or a large ball as the sample using some stochastic procedure, and *then* decide through another stochastic procedure which ball(s) should be placed in the urn. We can do that if we do not necessarily pick the sample from the urn. I would suggest that the persuasiveness of Bradley's argument hangs, in an illegitimate way, on this seeming silly. In this context, we should consider two analogies between Doomsday and urn scenarios. One is Bradley's, between his preferred analysis of Doomsday and a natural urn scenario. The other is between a different analysis of Doomsday, free of existential bias, and a silly urn scenario. We may be tricked into thinking that the naturalness of the urn scenario makes the former analogy a sound

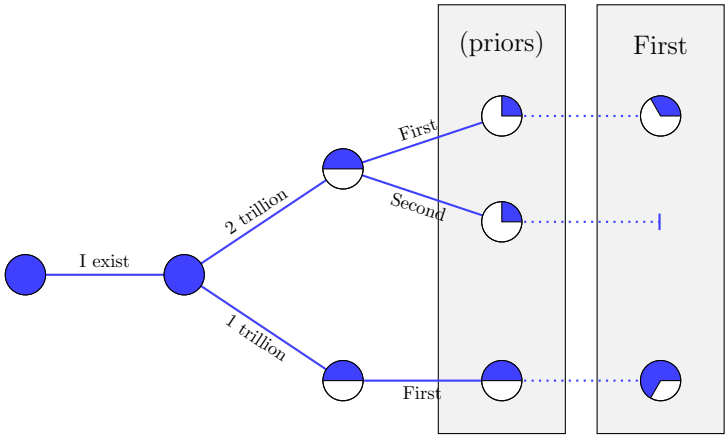

Figure 2 – Doomsday according to Halpern's first protocol and Bradley. It is here made explicit what the final probabilities are conditionalized on.

one—i.e., an analogy that licenses transfers of conclusions from one domain to the other—while the latter is not. But we shouldn't: the sample might not have been among the balls in the urn, and I might not have been among the existing people.

Sleeping Beauty: As noted above, Bradley's treatment of Sleeping Beauty is similar in structure to an urn scenario in which the urn contains one small and one large ball in the case of Heads, and two small balls in the case of Tails, where small balls represent Beauty being awake. Furthermore, he claims that it is a biased, rather than a random, selection procedure that is in play. The argument for this is quick (note that he uses the first person for Beauty):

> Was there a bias toward discovering a waking day rather than a sleeping day? That is, given the existence of a day on which I'm awake, is it certain that I observe a day on which I'm awake? Yes. If there is a day on which I'm woken, then I must observe a day on which I'm woken. So the case can be modeled using a procedure that is biased toward waking days. Bradley (2012, 169)

This reasoning contains a fallacy that I'm surprised Bradley didn't catch, given that immediately before this point in his argument, he discussed the difference between describing the evidence as "there is a day on which I am woken" and "I am woken today." The fallacy is based on an ambiguity in present-tense verbs: "I observe x" can (among other things) mean that the subject observes x *right now*, or that there is *some* instant of time at which the subject observes x. That it is "certain that I observe a day on

which I'm awake" is true in the latter sense, but not in the former. Beauty cannot be certain that she would have had the evidence of being awake *right now* under different circumstances, and thus the situation is not analogous to having the evidence of a small ball in a scenario where one could be certain of having exactly that evidence. There is no bias toward discovering a waking day rather than not discovering anything.

As in the case of Doomsday, we can analyze Bradley's assumptions in terms of a protocol. The "decision" about which day it is is assumed to come after (or depend upon) the coin toss: both Monday and Tuesday are genuine possibilities if and only if Tails was first decided, whereas "Nature" is forced to choose Monday if it first chooses Heads. Implicit in this assumption is another: that the very first decision was that Beauty must be awake. Here, it is helpful to look again at protocol 1 in Figure 1, but to consider explicitly what I left implicit then: an initial node connected to the first depicted node with a "Beauty is awake" link, similar to the "I exist" link in Figure 2.

So, according to Bradley, "Nature" first decided that Beauty must be awake, and then adjusted its decisions about the coin toss and the time to fit that premise.[5] That seems like a bad way to model the scenario. Surely Beauty's state depends on the coin toss and the time, not the other way around.

Quantum Mechanics: I will cover Bradley's treatment of Quantum Mechanics quickly, because the aspects of it that I am interested in commenting on are quite similar to the parallel aspects of Doomsday and Sleeping Beauty. On the one hand, my position is that Nature might have decided not to let me—i.e., the version of me that has just measured an Up spin—exist. If Nature had both decided on the stochastic version of quantum mechanics and on Down, then I would not have existed. *I* would not have been someone else instead. *My* existence is not a metaphysical necessity.

On the other hand, Bradley assumes that I must exist. In the analogy: *some* ball from the urn is drawn. In this case, a small ball represents Up and a large one represents Down. Heads and Tails represent the stochastic and the Everettian versions of quantum mechanics, respectively. Thus, the urn scenario claimed by Bradley to be structurally similar to Quantum Mechanics is the one in which Heads implies that the urn contains a small or a large ball with equal probability, and Tails implies that it contains one small and one large ball; and a ball is drawn using the random procedure. Hence, learning that I am Up does not confirm either the stochastic or the Everettian version of quantum mechanics (and neither would learning that I am Down). This analysis is illustrated in Figure 3.

Fine-Tuning: Bradley's treatment of Fine-Tuning at first seems inconsistent with his treatment of the three other cases. If I had had to guess how he

[5] Or perhaps more accurately: *failed* to adjust its decision about the coin toss. It still has a 50/50 probability distribution even though it is conditional on Beauty being awake.

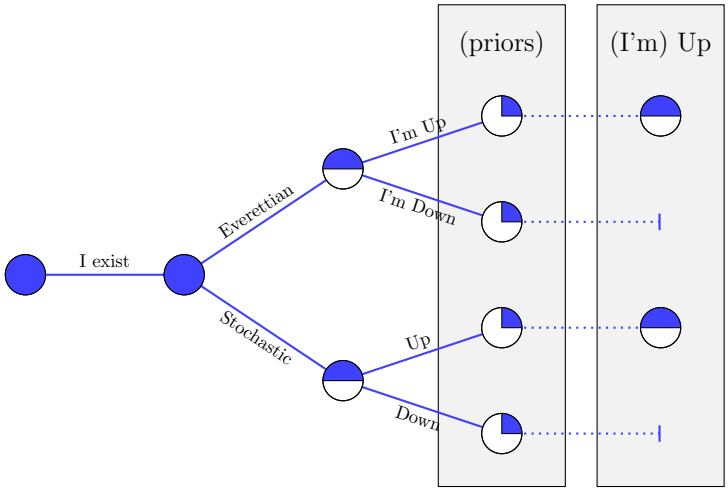

Figure 3 – Quantum Mechanics according to Bradley

would have analyzed Fine-Tuning, based on my knowledge of his analyses of the other cases, I would have suggested the following. Nature's first decision was for me to exist. Nature's next decision would be whether there should be one or two universes, and—unaffected by the first decision—it would be made stochastically, with uniform probability. Third, Nature would make a decision about the number of hospitable universes. Since my existence was already fixed, the probability of $\frac{1}{2}$ for one existing universe would pass on undivided to the possibility that there is one hospitable universe, while the probability of $\frac{1}{2}$ for two existing universes would be divided only between the options that there are one and two hospitable universes. The model would thus look like Figure 4, and my existence would not confirm anything.

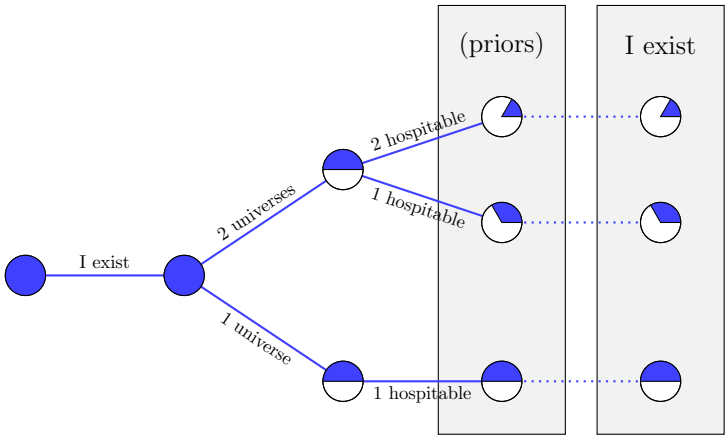

Figure 4 – Fine-Tuning according to pseudo-Bradley

However, that is not Bradley's analysis. In fact, he does *not* assume in this

case that I *must* exist. He accepts that, regardless of whether there are one or two universes, there is a chance that there is no hospitable universe containing people. Letting a small ball represent a hospitable universe, and a large one an inhospitable universe, the correct urn scenario is one in which Heads implies that the urn contains one small or one large ball with equal probability, and Tails implies that the urn contains two balls, each being small or large with equal and independent probability. That is, Heads represents the existence of one universe, and Tails, two.

The crucial assumption that Bradley *does* make is that in order to adequately model Fine-Tuning with an urn scenario, one should use the procedure biased in favor of small balls. That is, if there is a hospitable universe, I will live in it, and hence have the evidence that there is such a universe. So, while he does not assume categorically that I exist, Bradley does assume the hypothetical that if someone exists, then I exist.

In the language of "Nature" making "decisions," we thus have the following situation, according to Bradley. First, Nature decides between the existence of one and two universes with equal probability. Second, it decides for each existing universe whether it is hospitable with 50/50 probability for each universe, independently. Third, I am created if and only if there is a hospitable universe. Hence, my existence is more likely if there are two universes, so that hypothesis is confirmed. To be precise, conditional on my existence, the probability of there being two universes is $\frac{3}{5}$, as can be seen in Figure 5.

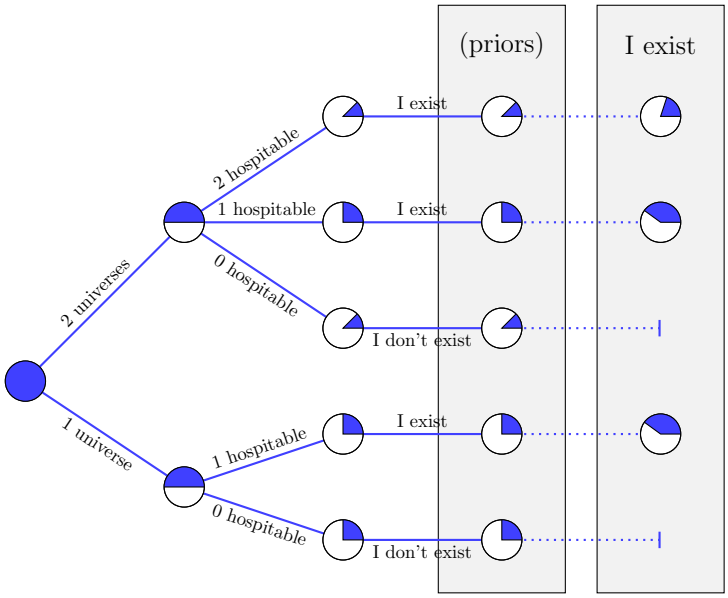

Figure 5 – Fine-Tuning according to Bradley

The existential bias implicit in Bradley's paper is thus not that I must exist,

but that I must exist if anyone does. It is as if Nature had placed my soul at the front of the line of potential people hoping to be actualized. It is very difficult, at best, to see how such an assumption of subjective privilege could be justified.[6]

4 Models without existential bias

My existence is, for me, *certain* in the epistemic sense of that word. That is, it is rational for me to assign it credence 1. Hence, it satisfies the condition that is required for me to conditionalize on it when updating other credences. However, I have no reason to think that my existence was certain in the modal sense of being *necessary*. Nor should I assume that I am modally privileged in some weaker way. The objective probability of my existence, as it was before I came into existence, should not be modeled differently from yours, or that of any other (potential) person.

In addition to *whether I am*, it is relevant when dealing with problems of self-location to ask *who I am* (Sleeping Beauty is slightly different, and I will get back to that). In this case, the opposite modal stance seems most plausible: I could not have been someone different from who I actually am, because that would presuppose that "I" and "who I actually am" refer to two different entities, like an immaterial soul and a physical body, respectively. However, who I am is in some situations *not* certain in the epistemic sense. And were it not for the empirical information that I have, I could not rule out any thesis about me being identical to any possible person, i.e., any thesis about which centered world is actual.[7] Hence, for epistemic purposes it is *as if* Nature had played dice with my identity and assigned my and other immaterial souls to (possible) physical bodies in a giant lottery. When assigning credences prior to all conditionalization on empirical information, I should pretend that it had.

Hence, I think that Halpern's metaphor of Nature making decisions about whether I exist, who I am, and what the world is like is entirely adequate for modeling purposes. We just disagree about the order of those decisions and, perhaps, their interpretation. That is, in spite of the language used by Halpern, I take the order to be not a temporal one, but one of dependency, i.e., one decision is "before" another iff the former influences the latter. If two decisions do not influence each other, they are "simultaneous."

How to model that order in a manner that avoids existential bias is quite clear, provided that we make the assumption—metaphysically implausible, but adequate for modeling—that Nature makes a decision about which souls

[6] While this is what he believes, he also shows that his conclusion follows from weaker assumptions with which I agree—see below.

[7] This is a point of agreement with Leslie (1990, 69).

to attach to which bodies. First, my existence must depend on what the world is like and which potential person my "soul" is assigned to, not the other way around. That is, I exist if and only if my soul is assigned to a potential body that exists pursuant to Nature's decision about what the world is like. Second, the decisions about the world and my identity must be independent, because if the decision about (among other things) whether person p exists increases or decreases the likelihood that I am person p (or vice versa), then there is existential bias. Hence, we should model as if my existence is determined by a probability distribution that is the product of the probability distribution for how the world is and the probability distribution for who I am.

So, to model the scenarios we are interested in, we just need to determine the latter two probability distributions. Regarding the how-the-world-is distribution, we will just continue to make use of simple example distributions, in line with the treatment in sections 2 and 3. And we assume that the modeled epistemic agent knows this distribution, and therefore must take it as his/her point of departure in line with the Principal Principle (Lewis 1980), as is also implicit in both Halpern's and Bradley's treatments. The decisive property of the who-I-am distribution, on the other hand, is dictated by the goal of not treating me as privileged: it has to be uniform. But uniform over which outcome space? Well, we can use the set of all possible people as it is according to the scenario, i.e., the union, over the outcome space for how the world is, of the set of people according to each outcome. We can do that if we choose the right subset of the agent's actual knowledge as the set of hypothetical knowledge that the hypothetical priors are based on. In Doomsday, we "remove" the knowledge that the agent is among the first trillion people; and in Quantum Mechanics, "remove" the knowledge that the Up event happened. In addition, in all the scenarios except Sleeping Beauty, we remove the knowledge that the agent actually exists, but not the knowledge that s/he *possibly* exists, i.e., not the knowledge that s/he exists in at least one of the possible worlds in the how-the-world-is outcome space.[8] The situation in Sleeping Beauty is similar, but with awakeness replacing existence.

Doomsday: The result of applying these principles to Doomsday is illustrated in Figure 6. The outcome space for how the world is consists of two options: that there will be one trillion, or two trillion, people in total; and we assume that each option has the same probability. The outcome space for who I am could have a cardinality of two trillion, but since my exact birth

[8] For comparison, imagine that you want to treat your actual knowledge that the outcome of a die roll is 6 as old evidence. One option is to consider the hypothetical priors under the assumption that you have *no* knowledge of the outcome, but it is not the only option. It would also be legitimate to consider the hypothetical priors under the assumption that you (merely) know that the outcome is even. This is similar to hypothetical priors under the assumption that you (merely) know your existence is possible.

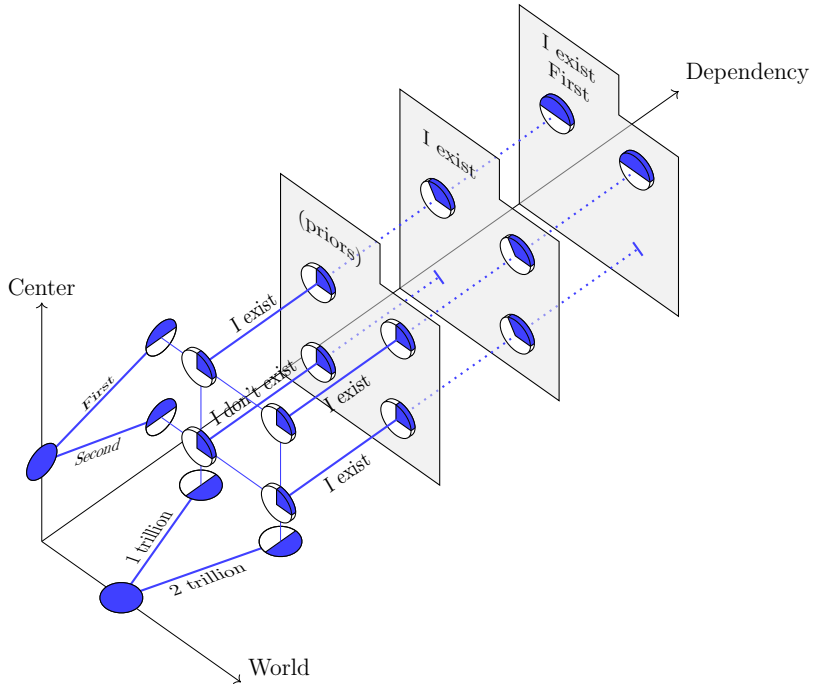

Figure 6 – Doomsday

rank doesn't matter, we can simplify in the way Halpern and Bradley do, and just let it contain two options—that I am in the first trillion, and that I am in the second trillion—with a probability of $\frac{1}{2}$ assigned to each. These two outcome spaces combine to form a product space with four options, each with a probability of $\frac{1}{4}$. Next, we bring in the variable that is my existence. Nature's decision about whether I exist is determined by the previous decisions, so the final outcome space also just contains four options, namely: (1 trillion, First, I exist), (1 trillion, Second, I don't exist), (2 trillion, First, I exist), and (2 trillion, Second, I exist). Their equal probability constitute the priors, i.e., the probabilities before I conditionalize on the subjective evidence that goes beyond my possible existence.

When I conditionalize on the fact that I exist, one of the options that involves the total human population being limited to one trillion people is eliminated. Hence, there is a probability shift away from the proposition that there will only be one trillion and towards the proposition that there will be two. My existence is not metaphysically necessary, and the fact that it happened anyway confirms the thesis that there are many people in the history of the universe.[9]

[9]This confirmation, it should be emphasized, is relative to an epistemic situation in which I only take objective evidence, plus the fact that my existence is possible, into account. We might, for instance, imagine that the objective evidence originates in the discovery of a natural process that once gave an objective probability of $\frac{1}{2}$ each to there

If I later learn that I am among the first trillion people, then another option is eliminated, leaving just two. The effect is that the original distribution of probability between the two contradictory theses about the total number of people is restored. Relative to the assessment made on the basis of non-subjective evidence, I cannot use my subjective evidence to predict the future (now that I know I'm not in it).[10]

Sleeping Beauty: The general considerations in the beginning of this section apply *mutandi mutandis* when time-slices of people are considered instead of people, and existence is replaced with the property of being awake. This allows us to model Sleeping Beauty in the same way as the three other scenarios. It is actually the simplest of the four to model: see Figure 7. The how-the-world-is outcome space consists of just Heads and Tails, and they have equal probability. The outcome space for *when* Beauty is only has to contain the times of the Monday awakening and the possible Tuesday awakening, if we let the priors be based on Beauty's knowledge that *now* is a possible awakening time during the experiment (while reserving the

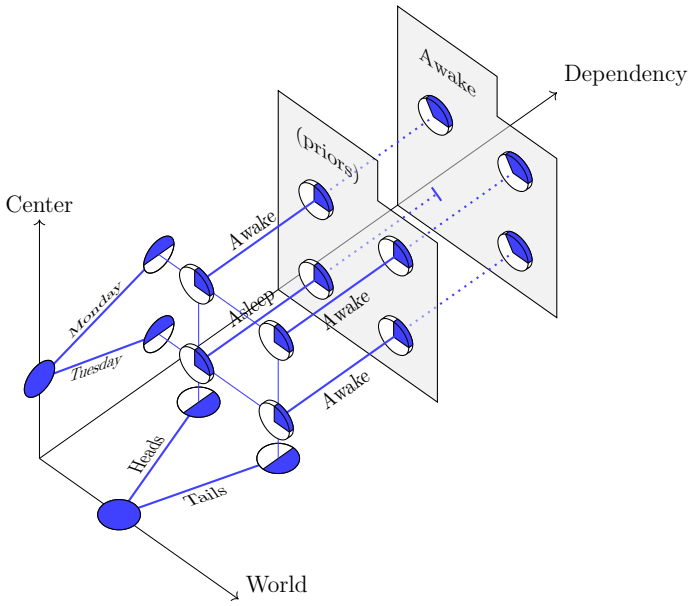

Figure 7 – Sleeping Beauty

being one trillion people and two trillion. It is thus *not* correct that treating my existence as non-trivial evidence implies by itself that these two hypotheses about the total number of people should be assigned probabilities in proportion to those two numbers, as Dieks (2007, 430) mistakenly claims. That only happens in the present example because the priors are equal. The general rule is that the priors should be multiplied by factors that are proportional to the total numbers of people predicted by the hypothesis.

[10]The result is thus the same as if one applies the so-called self-indication assumption in addition to the so-called self-sampling assumption. See Olum (2002) and references therein. I discuss the main objection to the former principle in section 6.

knowledge that *now* is an actual awakening time for conditionalization). And we assign a probability of $\frac{1}{2}$ to each, because Nature's "decision" about what time it is is not affected by Beauty's circumstances. Like in Doomsday, the two outcome spaces combine to form a product space with four options, each with a probability of $\frac{1}{4}$, and the space doesn't grow when we also take into account the variable for whether Beauty is awake or asleep, since that variable's value is determined by the values of the other two. When Beauty conditionalizes on the evidence that she is awake, the result are probabilities of $\frac{2}{3}$ for Tails and $\frac{1}{3}$ for Heads.[11] It is worth noting that if conditionalization on Awake&Monday were added at the back of Figure 7, figures 6 and 7 would be isomorphic.

Quantum Mechanics: However, the model for Quantum Mechanics looks quite different, as a glance at Figure 8 will reveal. Nature decides—with even probability, we will assume for the sake of example—whether quantum mechanics should be Everettian or stochastic. If it chooses the latter, it

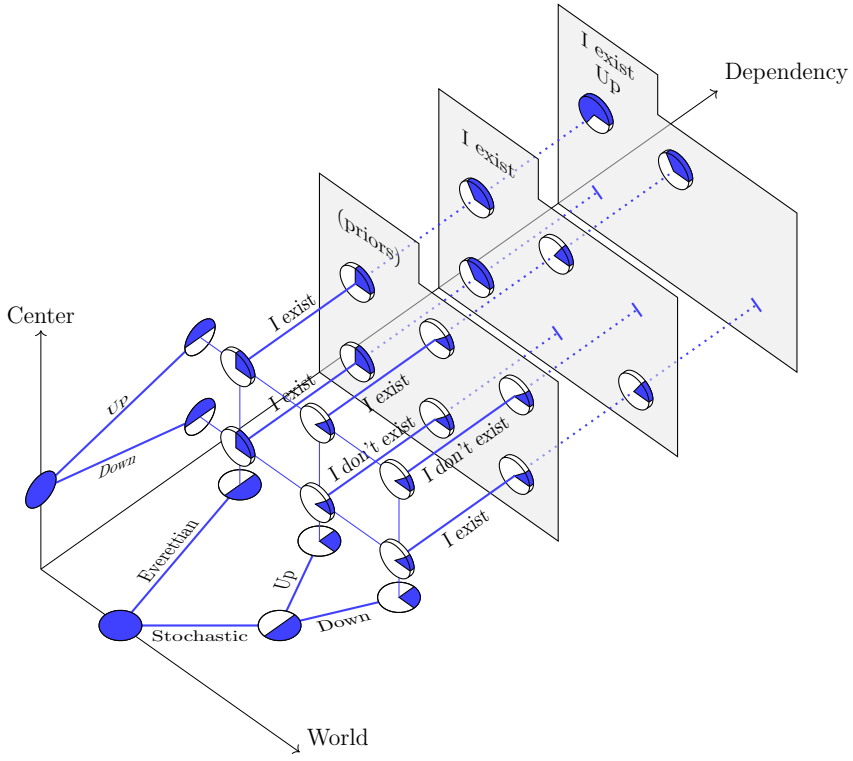

Figure 8 – Quantum Mechanics

[11] We have thus arrived at the solution to the Sleeping Beauty paradox formulated by Horgan (2004). See also the extensive subsequent exchange between Horgan and Mahtina on one side and Pust on the other (Horgan 2007; Pust 2008; Horgan 2008; Horgan and Mahtani 2013; Pust 2013; Pust 2014).

subsequently chooses between the Up event and the Down event, while it makes no more relevant decisions about how the world is in case of the former: it just lets both events happen. And, independently of its decisions about the world, Nature "decides" whether I am the potential Up person or the potential Down person.

Then, my existence is determined by the previous decisions: if the world is Everettian, I exist irrespective of whether I'm the Up or the Down person; but if the world is stochastic, I exist if and only if Nature's decision about Up versus Down in the world matches its decision about who I am. The prior probability of my existence is thus only $\frac{3}{4}$, and unevenly distributed between the Everettian and the stochastic options. Hence, when I update on the subjective evidence that I do indeed exist, the Everettian theory is always confirmed. Moreover, there is no potential self-locating evidence in this scenario that can restore the prior probability distribution between the two theories, like me being among the first trillion people could in Doomsday, and being told that it is Monday could in Sleeping Beauty. If, for instance, I learn that it was an Up event, the Down options will be eliminated, but the balance between the Everettian and stochastic options remains the same as it was when I had only conditionalized on my existence.[12]

Fine-Tuning: This is the only one of the four cases in which I agree with Bradley about the conclusion: my existence confirms that there are multiple universes. However, I do not agree with his reasoning. Mine is illustrated in Figure 9. As before, we assume that Nature makes a decision between a universe, on the one hand, and a multiverse consisting of exactly two universes, on the other; and that those two options have the same objective probability. I will further assume that one of the universes that exists if the multiverse does is identical to the sole universe that exists in the other option. This is merely for convenience, and does not affect the conclusion. Let "Universe 1" denote that universe, and "Universe 2" the possible second universe. Independently, Nature decides whether I am a (possible) person in Universe 1 or Universe 2, and it does so with uniform probability.

These decisions jointly determine whether my universe exists. But any universe, including mine, may or may not be hospitable. For the purpose of the model, I will assume that each universe has a probability of $\frac{1}{2}$ of being hospitable. Of course, that probability should be assumed to be independent of whether the universe in question is mine, because to assume otherwise would be a manifestation of existential bias.

The probability we assign to my existence, conditional on the existence of my universe, must be positive; otherwise, I could not have existed. But

[12]The proposition that there are computer simulations with conscious beings (Bostrom 2003) similarly gets a boost—and hence, so does the proposition that *I* am in such a simulation.

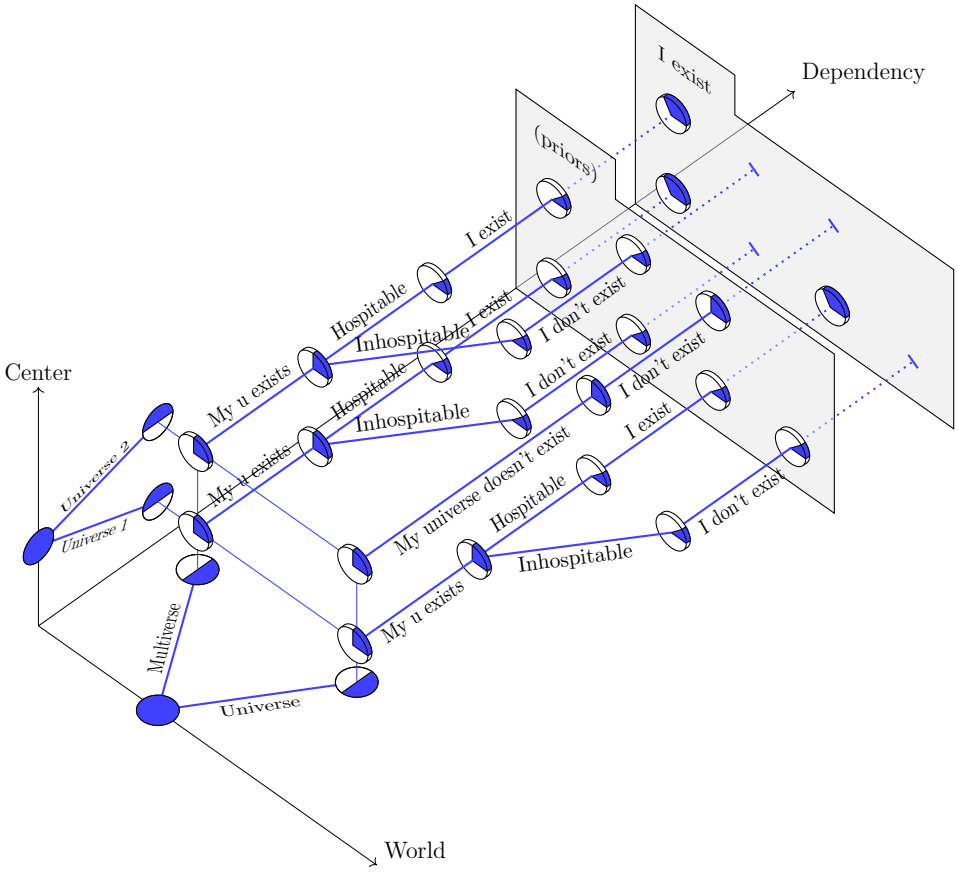

Figure 9 – Fine-Tuning

apart from that, the conclusion is unaffected by that probability, so we can just model it as being equal to 1. (This choice simplifies Figure 9.)

Our model does not have to take account of whether or not an existing universe that is not mine is hospitable. So, without that complication, we end up with seven possibilities. In three of them, with a combined probability of $\frac{3}{8}$, I exist. Of those three eighths, one eighth comes from the possibility that there is only one universe, and two eighths from the possibility of a multiverse. Hence, when I conditionalize on my existence, the multiverse theory is confirmed.

However, it should be noted that my agreement with Bradley is superficial. According to him (and many others, e.g., Leslie 1989, Parfit 1998, Smart 1989, and van Inwagen 2009, chapter 9), the multiverse makes it likely that I live in a hospitable universe, even though each universe is unlikely to be hospitable. But as a moment's reflection on Figure 9 reveals, my conclusion would have been the same if each universe's probability of being hospitable (conditional on its existence) were 1 instead of $\frac{1}{2}$. Therefore, in my account,

the role of the multiverse is not to explain why my universe is hospitable, and the confirmation of the multiverse hypothesis is not due to it making such hospitable-ness more likely. Rather, the multiverse just makes it more likely that my universe exists.[13] If it is very unlikely for a universe to be fine-tuned, then I accept that I am simply very lucky to be alive.

5 Hypothetical priors

One of the justifications I have given above for my theory is that, if you make use of hypothetical priors, it is what you arrive at. However, such priors are controversial. One potential problem, usually discussed in the context of scientific evidence and under the heading "the problem of old evidence," is that the hypothetical is a situation in which the epistemic agent is ignorant of a proposition, A, that they actually know; and there might be different potential ways for them not to know it, none of which is clearly maximally similar to the actual situation. For example, the agent might know A because it follows from the conjunction of two other known propositions, B and C, but not from either conjunct individually. Hence, it is not clear whether the hypothetical priors should be those the agent would have had if s/he were ignorant of B, or those the agent would have had if s/he were ignorant of C (Chihara 1987). Moreover, there is a second and more specific potential problem, namely that the hypothetical situation that is relevant for present purposes is one that rational epistemic agents cannot possibly be in: one in which they are ignorant of their own existence. And, allegedly, reliance on an impossible hypothetical is illegitimate (Draper, Draper, and Pust, see footnotes 11 and 13).

I believe the approach survives these challenges, and that Howson (1991) and Horgan and Mahtani (again, see footnote 11) have done an excellent job of showing that. However, for those of you who are not convinced, I will provide an alternative justification. It also relies on a hypothetical, but it avoids the alleged problems.

Imagine that at some point during your existence you are offered the opportunity to replace your credence function with that of another epistemic agent whose set of evidence includes the evidence you have at the time, and then update *that* system for the rest of your life. Would it be rational for you to accept the offer? Well, that depends. The other agent might have arrived at his credences through an irrational process. So let us stipulate that this did not happen, i.e., that he has abided by all rules of Bayesian

[13] My position was considered by Draper, Draper, and Pust (2007), who rejected it based on the premise that hypothetical priors must be ones that a rational agent can have in some situation. This requirement seems entirely ad hoc to me. We do not deny the evidential value of "I exist" for purposes of deduction, so why should we for purposes of probabilistic reasoning? Further discussion of this point can be found in the next section.

rationality (whatever they are, which I assume includes the Principal Principle). Even so, there might be another reason you would not want to adopt his credences: that the priors he started out with, whenever he came into existence, might be objectionable to you. But let us just stipulate our way out of this objection as well: his ultimate priors were some that you deem to be perfectly reasonable and rational for him to have started out with. With these amendments to the thought experiment, I believe the answer to the questions is clear. That is, you *should* accept the offer and adopt his credences as your own. From that premise, I can argue for my desired conclusion.

To do so, I need to fill in some more details about this other epistemic agent, while respecting the requirements that have been established so far: i.e., that his set of evidence includes yours, that he is rational, and that his ur-priors are what you would have preferred. The first detail is that he exists necessarily and knows that—perhaps not existed from the beginning of time, just from some point that precedes your creation.[14] This fact means that it is uncontroversial that *he* should not adjust his credences based on his own existence in the way I argue the rest of us should. Thus, if he knows the objective probabilities of the how-the-world-is propositions that are central to the scenarios in which we are interested, and has no other relevant evidence, his credences will match those probabilities. This is because he abides by the Principal Principle.

From the requirement that his set of evidence includes yours, this conditional follows: if you are created, he learns that. However, to be able to infer something interesting from this fact about him, we need to be careful about the protocol by which he learns about your existence. If, for example, the protocol were that he necessarily learns about the creation of exactly one person and that person was chosen at random, and just happened to be you, it would not be interesting for our purposes. Instead, let the protocol be: necessarily, if you are created, he learns about it; necessarily, the only other contingent individuals' existence he learns about are those *you* know about at the point in time where you are offered his credences; and he knows all this.

Before learning of your creation, what is his credence for that event, conditional on each of the two how-the-world-is propositions of a given scenario? Well, it had better be proportional to the number of subjects according to each proposition, because otherwise he considered you modally privileged (or underprivileged). And then, since you approve of his priors, you think that it is rational for other subjects to consider you so.

If it is proportional, he will, when he learns about your creation, adjust his

[14] In the case of Sleeping Beauty, he should be necessarily awake, and unable to retain the memory of Monday on Tuesday.

credences by factors that are proportional to the number of subjects in each possibility. Hence, if you were to adopt his credences, your credences would be similarly adjusted. And in the limit case of him only having the same evidence as you have, not more, your credences would be exactly what I argue they ought to be.

Admittedly, there is a slight wrinkle here. He must know about his own existence, and thus we have to assume you do too. So strictly speaking, this argument only applies to possible worlds in which you would know that kind of thing—and hence, where it would be true. However, I will put the burden on you to argue that there is a relevant difference such that the credence shift is appropriate in those possible worlds and not in ours. I believe this argument is convincing even though there is no actual necessarily existing agent with the described properties, just as I find Dutch Book arguments convincing concerning scenarios where no actual Dutch bookie is present.

This justification avoids the two problems mentioned at the beginning of this section: the first, because it is clear what is known in the hypothetical situation, and the second, because the initial ignorance of your existence is rationally possible, as it is someone else's ignorance.

However, this might not convince you if you reject, across the board, rational agents' ability to determine credences for hypothetical situations. In that case, unfortunately, you and I might simply have to agree to disagree.

6 Bullets to bite?

I have shown that the approach of modeling the processes by which the agents of the four scenarios acquire their evidence does not necessarily lead to Bradley and Halpern's conclusions; and that if the agents' own existence (or awakeness) is modeled as non-necessary, then the conclusions are almost the opposite of theirs.

So far, so good. Or, so bad, depending on what your intuitions are. If you are like the median philosopher, you will have welcomed my conclusions concerning Doomsday and Sleeping Beauty, but felt somewhat uneasy, or worse, about my conclusion in the case of Quantum Mechanics and *the reasons for* my conclusion in Fine-Tuning. That bad gut feeling may assert itself most unpleasantly when we consider a fifth case, courtesy of Bostrom (2002, 124).

The Presumptuous Philosopher: Assume that at some point in the future, the objective evidence available to the scientific community shows with certainty that there will be either a trillion trillion subjects in the history of the universe, or a *trillion* trillion trillion subjects; and the objective evidence does not favor one option over the other, so prima facie it would seem that

the rational credence for each is $\frac{1}{2}$. A presumptuous philosopher then claims that if each of us also takes into account the subjective evidence that he or she exists, we should consider it virtually certain that the larger number of total subjects is correct.

My line of reasoning supports this conclusion. Yet, the presumptuous philosopher cannot possibly be right! It is outrageous to think that a scientific dispute of this sort can be adjudicated from a philosopher's armchair! The track record of such a prioristic arrogance is one of aether, celestial orb cosmology, and atoms the shape of Platonic solids! It is preposterous, ludicrous, ridiculous, and, indeed, presumptuous! Something must be wrong!!!

Right?

I don't think so. In fact, that is exactly how I would reason if I were in the described epistemic situation.

To align my subjective credences to the probabilities prescribed by the objective evidence alone, I would have to not consider the fact that I exist as evidence for the 10^{36}-subjects hypothesis over the 10^{24}-subjects hypothesis. Doing that, in turn, requires that I assign the same prior probability to my existence conditional on each hypothesis. While I can consistently do so without violating any of the formal rules of Bayesianism, it would be most unreasonable. This is because, under the supposition (to which, *ex hypothesis*, I am assigning probability $\frac{1}{2}$) that there are 10^{36} subjects, I cannot similarly assign the same prior probability to the existence of subject s conditional on each hypothesis, for all s in the set of $10^{36} - 1$ subjects that are not me. That is, I must consider myself specially privileged. *That* is presumptuous![15]

Some of the most significant historical moments of intellectual progress have consisted of the subversion of deeply held prejudices about us being specially privileged. It turned out that our tribe was not located at the center of the planet, nor our planet at the center of the solar system, nor our solar system at the center of the universe. We further learned that our species was not fundamentally different from others, and is not the only one to enjoy intelligence, emotionality, morality, or even the ability to use language and tools. Similarly, we would be wise to rid ourselves of implicit assumptions of modal privilege. And thus relieved, we should shift our credences in favor of the multiverse and Everettian theses, and—unlike in the above toy examples involving just a two-universe multiverse and a single particle in

[15]This is essentially Olum's (2002) argument. However, even he is skittish about accepting the presumptuous philosopher's reasoning, and considers avoiding it by assuming a principle according to which how-the-world-is credences should be adjusted in inverse proportion to world size. It should also be remarked that Manley (2022) has demonstrated, with a slight change to the Presumptuous Philosopher scenario, that the principles of Bostrom and Bradley have equally "presumptuous" consequences.

up/down superposition—massively so.[16]

References

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly 53*, 243–255.

Bradley, D. (2011). Confirmation in a branching world: The Everett interpretation and Sleeping Beauty. *The British Journal for the Philosophy of Science 62*, 323–342.

Bradley, D. (2012). Four problems about self-locating belief. *Philosophical Review 121*, 149–177.

Carter, B. (1983). The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London 310*, 347–363.

Chihara, C. S. (1987). Some problems for Bayesian confirmation theory. *The British Journal for the Philosophy of Science 38*, 551–560.

Dieks, D. (2007). Reasoning about the future: Doom and Beauty. *Synthese 156*, 427–439.

Draper, K., P. Draper, and J. Pust (2007). Probabilistic arguments for multiple universes. *Pacific Philosophical Quarterly 88*, 288–307.

Eddington, A. (1939). *The Philosophy of Physical Science*. Cambridge University Press.

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

Friederich, S. (2021). *Multiverse Theories: A Philosophical Perspective*. Cambridge University Press.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.

Halpern, J. (2015). The role of the protocol in anthropic reasoning. *Ergo 2*, 195–206.

Horgan, T. (2004). Sleeping Beauty awakened: New odds at the dawn of the new day. *Analysis 64*, 10–21.

Horgan, T. (2007). Synchronic Bayesian updating and the generalized Sleeping Beauty problem. *Analysis 67*, 50–59.

[16]I am grateful for the help and feedback I have received from Simon Friederich, Yan Chunling, Zhai Chenyu, and the anonymous referee.

Horgan, T. (2008). Synchronic Bayesian updating and the Sleeping Beauty problem: Reply to Pust. *Synthese 160*, 155–159.

Horgan, T. and A. Mahtani (2013). Generalized conditionalization and the Sleeping Beauty problem. *Erkenntnis 78*, 333–351.

Howson, C. (1991). The 'old evidence' problem. *British Journal for the Philosophy of Science 42*, 547–555.

Leslie, J. (1989). *Universes*. Routledge.

Leslie, J. (1990). Is the end of the World nigh? *The Philosophical Quarterly 40*, 65–72.

Leslie, J. (1996). *The End of the World*. Routledge.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, Volume II. University of California Press.

Manley, D. (2022). On being a random sample. Unpublished manuscript.

Olum, K. (2002). The doomsday argument and the number of possible observers. *The Philosophical Quarterly 52*, 164–184.

Page, D. (1999). Can quantum cosmology give observational consequences of many-worlds quantum theory? In C. P. Burgess and R. Myers (Eds.), *General Relativity and Relativistic Astrophysics*, pp. 225–232.

Parfit, D. (1998). Why anything? Why this? *London Review of Books 20*. January 22.

Pust, J. (2008). Horgan on Sleeping Beauty. *Synthese 160*, 97–101.

Pust, J. (2013). Sleeping Beauty, evidential support and indexical knowledge: Reply to Horgan. *Synthese 190*, 1489–1501.

Pust, J. (2014). Beauty and generalized conditionalization: Reply to Horgan and Mahtani. *Erkenntnis 79*, 687–700.

Smart, J. J. (1989). *Our Place in the Universe: A Metaphysical Discussion*. Blackwell.

van Inwagen, P. (2009). *Metaphysics* (3rd ed.). Westview Press.