Simulation and self-location

Peter J. Lewis[[1]](#footnote-1)

Don Fallis[[2]](#footnote-2)

**Abstract**

It is possible that you are living in a simulation—that your world is computer-generated rather than physical. But how likely is this scenario? Bostrom and Chalmers each argue that it is moderately likely—neither very likely nor very unlikely. However, they adopt an unorthodox form of reasoning about self-location uncertainty. Our main contention here is that Bostrom’s and Chalmers’ premises, when combined with orthodoxy about self-location, yields instead the conclusion that you are almost certainly living in a simulation. We consider how this (surprising) conclusion might be resisted, and show that the analogy between Sleeping Beauty cases and simulation cases provides a new way of evaluating approaches to self-location uncertainty. In particular, we argue that some conditionalization-based approaches to self-location are problematically limited in their applicability.

**1. Simulation and self-location**

Given some assumptions about the nature of conscious experience and the nature of computation, it is possible that you are living in a *simulation*—that your world is computer-generated rather than physical. But how likely is this scenario? Bostrom (2003) and Chalmers (2022, p. 101) each argue that it is *moderately* likely—somewhere around 30%. One might well find this conclusion somewhat unsettling. However, both Bostrom and Chalmers adopt an unorthodox form of reasoning about self-location uncertainty. Our main contention here is that their premises, when combined with orthodoxy about self-location, yields instead the conclusion that you are *almost certainly* living in a simulation. This conclusion is likely to strike most people as *seriously* unsettling, or maybe simply *absurd*.

So, while one might be happy to live with Bostrom and Chalmers’ moderate probability, one might be inclined to resist the extreme probability that actually follows from their premises. However, we will argue that this is no easy task. The high probability follows from the assumption that it is not too unlikely that humans will produce simulated worlds, and from standard reasoning about self-location. Both of these are plausible and widely accepted. We explore more indirect ways of resisting the conclusion, and show that they fare no better. But as a side-effect of this exploration, we show that simulation cases provide a new way of distinguishing between ways to think about self-location uncertainty.

Let us start by reviewing Bostrom’s and Chalmers’ arguments concerning the probability that you are living in a simulation. Bostrom (2003) considers three possibilities: either humans will become extinct before they develop the relevant technology, or they will develop the technology but will not be interested in simulating civilizations like ours, or they will develop the technology and they *will* simulate civilizations like ours. Bostrom judges these possibilities to be equally likely. But he notes that if the third possibility is correct, it is *almost certain* that you are living in a simulation. The reason is that if humanity develops the means and the inclination to produce simulations of civilizations like ours, they will most likely produce *many* such simulations, so the chance that you are living in one of those simulations is high. Bostrom concludes that the chance that you are living in a simulation is about 1/3, which is higher than you might initially have thought.

Chalmers (2022, p. 101) reasons in a similar fashion. Chalmers calls hypotheses that rule out the simulation of civilizations like ours “sim blockers”. He judges that there is a less than 75% chance that there are any sim blockers. And if there are no sim blockers, he argues that it’s plausible that at least 99% of intelligent beings are simulated—again, because there will be many simulations—and hence that you are almost certainly simulated (2022, p. 85). Overall, then, the chance that you are simulated is at least (roughly) 25%—again, higher than you might initially have thought.

Our concern here is not with the precise values of these probabilities. We are more interested in the way in which self-location figures in the above arguments. Both Bostrom and Chalmers assume a form of reasoning like the following:

I don’t know whether I am located in unsimulated base reality or in a simulation: my experience does not distinguish between the two possibilities.[[3]](#footnote-3) If there are many simulations, there are many locations in which I am simulated, but only one in which I am not. Hence, distributing my credence uniformly over the possible locations, the chance that I am simulated is high.

We do not take issue with this form of reasoning.[[4]](#footnote-4) Rather, we wish to note a limitation on the scope of self-location reasoning that is tacitly endorsed by both Bostrom and Chalmers. Note that in addition to self-location probabilities—probabilities over locations within the world—there are also *ordinary* probabilities—probabilities over what the world is like. For example, Bostrom assumes that there is a 2/3 chance that the world does not contain simulations of civilizations like ours. Bostrom and Chalmers each assume that the probabilistic effects of reasoning about self-location do not affect the probabilities of the worlds themselves. That is, even though reasoning about self-location makes it overwhelming likely that you are simulated *given* that the world contains simulations, self-location reasoning does not affect the probability that you are in such a world.

This kind of quarantining of the effects of self-location reasoning was once taken for granted (Lewis, 1983, p. 149). But the Sleeping Beauty puzzle challenges this assumption, and the canonical solution violates it (Elga, 2000). So let us explore what the Sleeping Beauty puzzle tells us about the chance that you are simulated.

**2. Simulation and Sleeping Beauty**

Consider the standard Sleeping Beauty puzzle. Beauty will fall asleep on Sunday, and will be awakened briefly on Monday. After she falls asleep again, a fair coin will be tossed. If it comes up tails, her memory of the Monday waking will be erased, and she will be awakened again briefly on Tuesday. If it comes up heads, she will sleep through Tuesday. Beauty knows all this. When she wakes up, what credence should she attach to the coin landing heads?

w1

w2

L1

L2

Fig. 1: Two worlds, two locations

Elga (2000) reasons as follows. There are two possible worlds, the heads world (w1) and the tails world (w2), and in the tails world there are two possible locations Beauty might occupy, namely Monday (L1) and Tuesday (L2), as shown in figure 1. Given that she is in the tails world, she will be unable to distinguish between the two waking locations, so by a suitable principle of indifference, her credence that this is her first waking and that this is her second waking should be the same. That is, P(L1|w2) = P(L2|w2). And since P(w2 & L1) = P(w2)P(L1|w2), and P(w2 & L2) = P(w2)P(L2|w2), this yields P(w2 & L1) = P(w2 & L2). Furthermore, given that this is her first waking, the probability of heads and tails are each 1/2, since these are probabilities for a future toss of a fair coin. That is, P(w1|L1) = P(w2|L1). And since P(w1 & L1) = P(L1)P(w1|L1), and P(w2 & L1) = P(L1)P(w2|L1), this yields P(w1 &L1) = P(w2 & L1). So we now have P(w1 & L1) = P(w2 & L1) = P(w2 & L2), and since these are exhaustive and mutually exclusive possibilities, they should each be assigned a credence of 1/3. And since P(w1 & L1) = P(w1)P(L1|w1), and P(L1|w1) = 1, we obtain P(w1) = 1/3. When Beauty wakes up, her credence that the coin lands heads should be 1/3. Given its reliance on indifference principles, we will call Elga’s approach the *indifference approach* to self-location.

Compare this with reasoning about whether you are living in a simulation. To begin with, let us consider a rather implausible simulation hypothesis, constructed so as to make it as closely analogous to the Sleeping Beauty case as possible: we can adjust it to make it more plausible later. That is, suppose (rather implausibly) that the world contains exactly one simulation. And suppose (even more implausibly) that the contents of that simulation are determined by a coin-toss in base reality. If the coin lands tails, the simulation simulates our world and its inhabitants. If the coin lands heads, the simulation is similar, except that all (potentially) conscious beings are in suspended animation. The structure of this example is exactly the same as Sleeping Beauty: there are two possible worlds, world w1 (the heads world) and world w2 (the tails world), and in the tails world, there are two possible self-locations, location L1 (base reality) and location L2 (simulated reality), again as shown in figure 1. Exactly the same reasoning can be applied. Given that you are in the tails world, you will be unable to distinguish between location L1 and location L2, so appealing to the same indifference principle, P(L1|w2) = P(L2|w2). Given that you are in base reality (location L1), the probability of heads and tails are each 1/2, since it is assumed that a fair coin is tossed in base reality. Hence P(w1|L1) = P(w2|L1). Putting these together yields a probability of heads of 1/3, as before.

Notice that in each case, the self-location uncertainty in one of the two worlds gives rise to a shift in the probabilities of the *worlds* themselves. Hence the assumption made by both Bostrom and Chalmers that the probabilities assigned to worlds are unaffected by self-location uncertainty is highly questionable. In the particular simulation hypothesis just described, the chance that a world contains (conscious) simulated people is 1/2, and given that you are in such a world, the chance that you are simulated is 1/2. That is, according to the reasoning endorsed by Bostrom and by Chalmers, the probability that you are simulated is 1/4. But if we pay attention to the analogy with Sleeping Beauty, it looks like the correct probability that you are simulated is the probability assigned to location L2, which is 1/3.

Finally, note that in the simulation case, the self-location uncertainty is *permanent* in a way that it is not in Sleeping Beauty. In the Sleeping Beauty case, when Beauty wakes up, she can ask what day it is: if the answer is “Monday”, her credence in heads reverts to 1/2. But the inhabitants of a simulation have no way of finding out that they are in a simulation: in the particular case described here, your credence in heads *remains* 1/3.

**3. Standard simulation scenarios**

The particular simulation hypothesis described so far is not at all plausible—and it doesn’t yield a particularly high probability that you are simulated. Let’s see what happens if we make the example closer to the standard scenario assumed in the literature. First, let’s assume, as Bostrom and Chalmers do, that if a world has the means and inclination to produce simulations of civilizations like ours, it will produce many of them. For the moment, we will continue the fiction that if a coin-toss in base reality comes up heads, all the inhabitants of the simulations are in suspended animation. That is, the only difference we are considering so far is that the number of possible self-locations in the tails world increases: there are now N such locations, of which only one is base reality, as shown in figure 2. Application of the indifference approach goes just as before. Given that you are in the tails world, you are unable to tell which location you occupy, so by indifference, the credences attached to each of the locations (w2 & L1), (w2 & L2), … (w2 & LN) are the same. Given that you are in base reality, the probability of heads and tails is 1/2 each, so the credences attached to (w1 & L1) and (w2 & L1) are the same. So you have the same credence in each of N + 1 exhaustive and mutually exclusive possibilities: your credence in each of them is 1/(N + 1). Hence your credence that you are in the heads world is 1/(N + 1). More importantly, your credence that you are simulated (i.e. not in location L1) is (N – 1)/(N + 1). As N gets large, this probability gets close to 1.

w1

L1

L2

w2

L3

LN

…

…

Fig. 2: Two worlds, N locations

Second, let’s remove the suspended animation element in the example: if the coin lands heads, no simulations are created at all, as shown in figure 3. The suspended animation element made the original simulation example more closely analogous to the Sleeping Beauty case, but note that it plays no role in the probabilistic reasoning. There are still two possibilities given that you are in base reality, heads and tails, with the same probability. And there are still N possible self-locations given tails, and since they are subjectively indistinguishable, they each get the same probability. Overall, then, there are N + 1 equiprobable possibilities, of which one is a heads location, yielding a credence in heads of 1/(N + 1), and a credence that you are simulated of (N – 1)/(N + 1).

w1

L1

L2

L3

LN

w2

…

Fig. 3: Two worlds, N locations in w2

Finally, let’s replace the coin toss with natural contingency. That is, there’s a chance that the world will contain no simulations: maybe we’ll die out before we develop the relevant technology, or maybe we’ll have no interest in simulating civilizations like ours. Suppose you judge that this chance is 1/2: then the reasoning goes exactly as before. The N possible self-locations on tails are equiprobable. And given that you are in base reality, the chance that you are in the world without simulations (w1) is the same as the chance that you are in the world with simulations (w2). That is, you judge the chance that we will develop and use simulations *under the assumption* that you occupy base reality, so you judge that P(w1|L1) = P(w2|L1). The result is the same: there are N + 1 equiprobable, exhaustive, mutually exclusive possibilities, in N – 1 of which you are simulated.

What if you judge the chance that we will develop and use simulations to be different from 1/2, say q? You still make that judgement under the assumption that you occupy base reality, so the only difference is that now P(w2|L1) = q and P(w1|L1) = 1 – q. Since P(w1 & L1) = P(L1)P(w1|L1), and P(w2 & L1) = P(L1)P(w2 |L1), we have P(w1 & L1)/P(w2 & L1) = (1 – q)/q. As before, P(w2 & L1) = P(w2 & L2) = … = P(w2 & LN). The result is N + 1 exhaustive, mutually exclusive possibilities, N of which have the same probability, and one of which has a probability (1 – q)/q as large. Hence the probability that there are no sims is (1 – q)/(Nq + 1 – q). Note that this probability gets arbitrarily small as N gets large. And the probability that you are simulated is (N – 1)q/(Nq +1 – q), or (N – 1)q/((N – 1)q + 1): note that this gets arbitrarily close to 1 as N gets large.

Let us return to the scenarios described by Bostrom and Chalmers. They estimate the chance that we will develop and use simulations to be 1/3 and 1/4 respectively. Bostrom claims that if there are simulations, there are likely to be many of them: Chalmers specifies that there are likely to be at least 99. Using Chalmers’ numbers, we have q = 1/4 and N = 100, so that the chance that you are simulated is 96%. Furthermore, Chalmers’ estimate is surely quite conservative: if computing power is cheap, then there could well be many thousands of simulations. Even if you judge that the chance that a world will develop and use simulations is relatively small, the result of self-location reasoning is that the chance that you are living in a simulation is close to 1.

**4. The structure of the argument**

We have argued that Bostrom and Chalmers’ premises, when combined with standard reasoning about self-location, lead to the conclusion that you are almost certainly living in a simulation. But note that we are not directly *defending* this conclusion. Rather, we are arguing for a conditional claim: *if* there is a non-tiny chance that we will produce many simulations, *then* you are almost certainly living in a simulation. This conditional is not a novel claim: sometimes it has been taken for granted,[[5]](#footnote-5) and sometimes a defense of it in terms of Sleeping Beauty has even been alluded to.[[6]](#footnote-6) But the power of this defense has not been fully appreciated, we think. Our goal here is to make the analogy with Sleeping Beauty fully explicit and assess its force.

Note, in particular, that we are not taking a position on how this conditional should be used: one might take the consequent to be *absurd*, and hence regard the appropriate inference as a *modus tollens* rather than a *modus ponens*. For most people (other than Elon Musk[[7]](#footnote-7)), the idea that you are almost certainly living in a simulation probably seems totally implausible. Furthermore, *if* you take the conclusion to be absurd, then there is further work to do in diagnosing the problem with the argument. Perhaps you should simply reject the premise that there is a non-tiny chance that humans will develop the appropriate kind of simulation. This is an empirical claim about the current trajectory and eventual limits of computing technology. We have nothing to add here to the assessments of others: see Bostrom (2003), Chalmers (2022, pp. 86–90), and Carlsmith (2022) for discussion. We note, though, that this claim seems to be widely accepted among philosophers. At the very least, it seems somewhat presumptuous to reject it on the basis of philosophical distaste for where it leads.

So perhaps, instead, you should reject one of the steps that leads from this premise to the conclusion that you are almost certainly living in a simulation. There are five such steps:

1. The argument that Beauty’s credence in heads on waking should be 1/3.
2. The analogy between the Sleeping Beauty puzzle and a simple simulation scenario.
3. The result of adding more simulations to this scenario.
4. The result of removing the unexperienced locations from this scenario.
5. The result of replacing the coin toss with natural contingency.

We consider the prospects for objecting to each step in the following section. Our contention is that none of them provides a straightforward way to avoid the problematic conclusion that you are almost certainly living in a simulation.

**5. Ways to object**

**5.1 Solutions to Sleeping Beauty**

The first step in our argument employs Elga’s indifference approach to conclude that Sleeping Beauty’s credence in heads on waking should be 1/3. But Elga’s conclusion has been challenged: some people argue that Beauty’s credence in heads on waking should be 1/2 (e.g. Lewis, 2001; Bostrom, 2007; Meacham, 2008; Bradley, 2011; Pust, 2012). If this is right, then the effects of self-location reasoning might, after all, be quarantined within a world, as Bostrom and Chalmers contend. We do not wish to wade into this extensive and complicated debate here. However, we note that “halfer” solutions are relatively unpopular (Bourget and Chalmers, 2023), and have generally been rebutted in the literature (e.g. Neal, 2006; Titelbaum, 2012).[[8]](#footnote-8) Anyone who wishes to object to our argument by adopting a halfer solution takes on the burden of defending it. At the very least, *if* you accept a thirder solution to Sleeping Beauty (and many people do), then you cannot object to the first step in our argument.

The situation regarding thirder solutions is more complicated. There are many people who argue that Beauty’s credence in heads should indeed be 1/3, but that the correct way to get to that conclusion is via a different approach from Elga’s (e.g. Neal, 2006; Briggs, 2010; Horgan and Mahtani, 2013; Milano, 2022; to name but a few). Insofar as they reach the same conclusion as Elga, such arguments do not challenge the first step in our argument. But insofar as they reach that conclusion via different reasoning methods, they raise the possibility that there is a *relevant* *difference* between the Sleeping Beauty case and simulation cases, so that the thirder conclusion does not carry over to the latter. We turn to such possibilities now.

**5.2 Parts and people**

The second step in our argument relies on the close analogy between the Sleeping Beauty case and a simple simulation case. But no analogy is perfect, and there are of course differences between the Sleeping Beauty case and the simple simulation case. Most pertinently, the Sleeping Beauty case involves the credences of a single individual at two times, whereas the simple simulation case involves the credences of two distinct individuals, one in base reality and one in simulated reality. One might worry that this disanalogy means that the 1/3 conclusion no longer follows in the simulation case.

Note that the indifference approach is immune from this worry: it relies only on the fact that Beauty *can’t tell* whether it is Monday or Tuesday. Since you can’t tell whether you are in base reality or in a simulation, the indifference principle applies equally to both cases. But some solutions to the Sleeping Beauty puzzle might seem to appeal to the temporal nature of the case. For example, Horgan and Mahtani (2013) and Milano (2022) argue that Beauty should *update* her credences based on new information she obtains on waking. Let us call such approaches *updating approaches*. Updating approaches might be thought to rely on the assumption that there is a single individual existing across the various locations. So if *this* is the right way to reason about the Sleeping Beauty case, then perhaps it doesn’t apply to the simulation case.

However, it is worth noting that thirder reasoning is often appealed to in cases that involve distinct people rather than distinct temporal parts of a single person—for example cases involving duplicates (Elga, 2004),[[9]](#footnote-9) Boltzmann brains (Kotzen, 2020), or Everettian branches (Groisman, Hallakoun, and Vaidman, 2013). That is, it is generally assumed that thirder reasoning *does* carry over from the Sleeping Beauty case to cases involving distinct people. Admittedly, such cases are sometimes intended to be *reductio* arguments against thirder reasoning. For example, in his Presumptuous Philosopher example, Bostrom (2007, p. 64) argues that thirder reasoning entails that a physical theory in which there are many observers is more probable than a theory in which there are fewer observers, simply in virtue of the number of possible self-locations in the former. Bostrom finds this conclusion absurd, and uses it against thirder reasoning.[[10]](#footnote-10) But crucially, Bostrom’s argument relies on *assuming* that thirder reasoning applies equally to cases involving distinct people.

Indeed, as Neal (2006, p. 17) points out, even in the Sleeping Beauty case, the temporal element seems inessential: the only information that Beauty need retain between Sunday and her later awakenings is the set-up of the experiment, so Beauty need not be interpreted as really *updating* her earlier credences. Sometimes, admittedly, updating *is* interpreted in this literal way: Milano (2022) describes Beauty as calculating, on *Sunday*, conditional credences (such as the probability it’s Tuesday given that she’s awake), and then using those conditional credences, on *waking*, as the basis of her unconditional credence that it’s Tuesday. But there is no need for her to proceed this way: she could equally generate the conditional credences *on waking* by temporarily bracketing the information she has that she is awake (Horgan and Mahtani, 2013). The latter procedure carries over straightforwardly to cases involving distinct *individuals*, such as the simple simulation scenario: each person in such a scenario can calculate conditional credences (such as the probability that I am living in a simulation given that I’m not in suspended animation), and then use those conditional credences as the basis of the unconditional credence that I am living in a simulation. It makes no difference whether the credences concern which *day* it is or which *person* I am. That is, Milano’s approach, taken literally, is artificially limited to cases with a *temporal* possibility structure, but there is no need for updating approaches in general to be so limited.[[11]](#footnote-11)

To drive last this point home, consider the following case (adapted from Neal’s (2006, p. 17) Sailor’s Child problem).[[12]](#footnote-12) An infertile couple tries to conceive a child via IVF and surrogacy. The fertility clinic case has two embryos from the couple in their freezer, each labelled with its time of creation. The fertility clinic director decides how many embryos to implant by tossing a fair coin: if it lands tails, two embryos are implanted in two surrogates, one embryo in each; if it lands heads, the first-created embryo is implanted in a single surrogate, and the second-created embryo is destroyed. Eight months later, the infertile couple are killed in a car accident, and under the terms of the surrogacy contract, any children born through surrogacy are adopted. You know that you are an adopted child, and a little digging into the circumstances of your birth uncovers this story. What is the probability that you have a sibling—i.e. that the coin came up tails?

It seems hard to resist the conclusion that the answer is 2/3, even though this case involves two *people* on tails rather than two temporal parts of the *same* person. This straightforwardly follows under an indifference approach: given tails, you can’t tell whether your embryo was first or second. But even under an updating approach, the same conclusion follows. There are two independent random variables in this case—heads or tails for the coin-toss, first or second for the embryo—yielding four possibilities with probabilities of 1/4 each. But one of these possibilities is inconsistent with your experience: you can rule out “heads and second,” resulting in three possibilities with probabilities of 1/3 each. This kind of belief updating doesn’t require that there is a single person involved.

**5.3 More locations**

The third step in our argument involves adding more simulations to the simple scenario. Once you have accepted that thirder reasoning applies to the simple simulation scenario, then adding more simulations looks like a trivial change. Indeed, one can add more waking locations to the Sleeping Beauty case as well, to the same effect (e.g. Bostrom 2007, p. 62).[[13]](#footnote-13)

**5.4 Missing locations**

The fourth step in our argument involves removing the unexperienced locations, replacing worlds in which all the simulated conscious beings are in suspended animation with worlds in which there are no simulations at all. Again, this makes no difference under an indifference approach: the approach assumes only that (w1 & L1), (w2 & L1), and (w2 & L2) are exhaustive locations of experience, and it makes no difference whether there are further locations that are not experienced, or no further locations at all.

But updating approaches to the Sleeping Beauty puzzle make explicit use of the fourth location (e.g. Horgan and Mahtani 2013, p. 339; Milano 2022, p. 661). These solutions begin by motivating non-zero credences for all four locations, and then proceed by ruling out the “heads and Tuesday” (w1 & L2) location and updating credences in the remaining locations accordingly. One way to make this approach vivid is via Neal’s (2006) inclusion of a second character, the Prince, who is treated exactly like Sleeping Beauty except that he is awakened on Monday and on Tuesday regardless of the outcome of the coin toss.[[14]](#footnote-14) The Prince has an initial credence on waking of 1/4 in each of the four possible locations, but when he sees that Beauty is awake, he conditionalizes on the fact that he is not in the heads and Tuesday location, resulting in a credence of 1/3 in the remaining three locations. Beauty can ask the Prince for his credences, and adopt them as her own: the Prince has exactly the same information that Beauty has. But Beauty doesn’t have to actually *ask*, because she can figure out for herself what credences he must have. In fact, the Prince doesn’t even need to *exist*; the fact that he *would* have these credences is good enough (Neal 2006, p. 16). The strategies of Horgan and Mahtani (2013) and Milano (2022) can be regarded as alternative accounts of Beauty’s internal monolog as she convinces herself that, since there are four locations with initial probability 1/4 each, and this can’t be location (w1 & L2), each of the remaining three has a probability of 1/3.

Such accounts, though, rely on the existence of the (w1 & L2) location as a *possible* location—and the fourth step in our argument involves *removing* such locations. You can’t obtain the low probability of heads in such a case by eliminating the possibility that the coin landed heads and you are simulated, because no such possibility exists. It isn’t that updating accounts suggest a *different* probability from indifference accounts in such cases: they are simply *inapplicable*, and hence do not suggest any probability at all. Even so, it looks like the fourth step in the argument creates a relevant difference between Sleeping Beauty and simulation, at least for *some* ways of thinking about self-location.

To gain some clarity here, let us return the IVF case. Recall that in the original version, the fertility clinic has two embryos in their freezer, each labelled with its time of creation: if the coin lands tails, both embryos are implanted, but if the coin land heads, the first-created embryo is implanted and the second-created embryo is destroyed. This provides the four possibilities that ground the updating approach. But what if the embryos are created on demand? That is, if the coin lands tails, two embryos are created—first one, then the second—and if the coin lands heads, one embryo is created. Surely it can’t make any different to your reasoning whether or not one extra embryo was created and destroyed, so the conclusion that the probability of heads is 1/3 should follow as before. Updating approaches that involve eliminating a possibility cannot reach this conclusion: there is no possibility to eliminate. But we know what the right answer should be in this case. Hence updating approaches seem unduly limited in their applicability: they require a possibility structure that happens to be present in Sleeping Beauty, but is in fact irrelevant to the nature of the reasoning.[[15]](#footnote-15)

Finally, note that the creation-on-demand variant of the IVF case is closely analogous to a simulation case in which there is a single simulation on tails and no simulation at all on heads. The analogy maps the first-created embryo to base reality, and the second-created embryo to the simulated world. What the IVF case suggests is that it really doesn’t matter whether “heads and simulation” is an unexperienced possibility (with all the inhabitants of the simulation in suspended animation), or no possibility at all (because no simulation was created). If the thirder conclusion follows in the former case (which it surely does), then it follows equally in the latter case. Further, the IVF case provides a new way of *evaluating* updating approaches: the fact that they don’t generate the thirder conclusion in the latter case is a *limitation*. The indifference approach, and indeed other approaches,[[16]](#footnote-16) apply seamlessly to the IVF case and to simulation cases: this should be taken as an argument in their favor. And these approaches lead, under standard assumptions, to the conclusion that you are almost certainly living in a simulation.

**5.5 The final step**

The final step in the argument involves replacing the coin toss with natural contingency. But presumably nobody thinks that there is anything special about coin tosses: the last step is innocuous. Hence the initial premise of the argument that you are almost certainly living in a simulation, and the five steps that leads from that premise to the conclusion, are all prima facie defensible. If the conclusion is absurd—and it may be—then diagnosing the problem with the argument is not a straightforward task.

**6. Conclusion**

What, then, is the probability that you are living in a simulation? If we accept Bostrom and Chalmers’ assumption about the chance that we will develop simulation technology, and accept standard reasoning about self-location, it follows that you are *almost certainly* living in a simulation. It is not, as Bostrom and Chalmers contend, only moderately likely: such a conclusion depends on an unwarranted “quarantining” of self-location uncertainty within a world.

This is a hard conclusion to live with, and so gives us added impetus to investigate the reasoning that gets us there. But there is no attractive strategy for resisting the argument. The initial premise is plausible and widely-held, and the steps in the argument from the premise to the conclusion are all defensible via prominent strategies for reasoning about self-location. Updating approaches to self-location might seem to offer some hope here, since they distinguish between the Sleeping Beauty case and simulation cases. However, this distinction just reveals a limitation in updating approaches: their applicability depends on irrelevant features of the possibility structure. This last point, we think, is independently interesting: comparing Sleeping Beauty cases to simulation cases reveals a way of evaluating updating approaches that is unavailable if one concentrates on Sleeping Beauty cases alone.

**Acknowledgements**

The authors are grateful to Dave Chalmers, Jonathan Fiat, Kay Mathiesen, and anonymous referees for this journal, for comments that improved this paper immensely.

**References**

Birch, Jonathan (2013). On the ‘simulation argument’ and selective scepticism. *Erkenntnis*, 78, 95-107. https://doi.org/10.1007/s10670-012-9400-9

Bostrom, Nick (2003). Are we living in a computer simulation? *Philosophical Quarterly*, 53, 243-55. https://doi.org/10.1111/1467-9213.00309

Bostrom, Nick (2007). Sleeping Beauty and self-location: A hybrid model. *Synthese*, 157, 59-78. https://doi.org/10.1007/s11229-006-9010-7

Bourget, David, and David J. Chalmers (2023). Philosophers on philosophy: the 2020 PhilPapers Survey. *Philosophers’ Imprint*, 23, 1-53. https://doi.org/10.3998/phimp.2109

Bradley, Darren J. (2011). Self-location is no problem for conditionalization. *Synthese*, 182, 393-411. https://doi.org/10.1007/s11229-010-9748-9

Briggs, R. A. (2010). Putting a value on Beauty. T. S. Gendler, and J. Hawthorne (eds.), *Oxford Studies in Epistemology, Volume 3.* Oxford University Press, 3-34.

Carlsmith, Joe. (2022). Simulation arguments. *Less Wrong*, 18 February 2022. https://www.lesswrong.com/posts/X7vdn4ANkdNwoSyxB/simulation-arguments

Chalmers, David J. (2022). *Reality+*. W. W. Norton.

Elga, Adam (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60: 143-47. https://doi.org/10.1093/analys/60.2.143

Elga, Adam (2004). Defeating Dr. Evil with self‐locating belief. *Philosophy and Phenomenological Research*, 69, 383-396. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x

Groisman, Berry, Na’ama Hallakoun, and Lev Vaidman (2013). The measure of existence of a quantum world and the Sleeping Beauty problem. *Analysis*, 73, 695-706. https://doi.org/10.1093/analys/ant072

Hansen, Casper Storm (2023). Existential bias. *Episteme*, 20, 701-721. https://doi.org/10.1017/epi.2022.57

Horgan, Terry, and Anna Mahtani (2013). Generalized conditionalization and the Sleeping Beauty problem. *Erkenntnis*, 78, 333-351. https://doi.org/10.1007/s10670-011-9316-9

Kotzen, Matthew (2020). What follows from the possibility of Boltzmann brains? S. Dasgupta, R. Dotan, and B. Weslake (eds.), *Current Controversies in Philosophy of Science.* Routledge, 21-34.

Lewis, David K. (1983). Attitudes *de dicto* and *de se*. *Philosophical Papers, Vol. 1*. Oxford University Press, 133-159.

Lewis, David K. (2001). Sleeping Beauty: reply to Elga. *Analysis*, 61, 171-176. https://doi.org/10.1111/1467-8284.00291

Lewis, Peter J. (2013). The doomsday argument and the simulation argument. *Synthese*, 190, 4009-22. https://doi.org/10.1007/s11229-013-0245-9

Meacham, Christopher J. G. (2008). Sleeping Beauty and the dynamics of de se beliefs. *Philosophical Studies*, 138, 245-269. https://doi.org/10.1007/s11098-006-9036-1

Milano, Silvia (2022). Bayesian Beauty. *Erkenntnis* , 87, 657-676. https://doi.org/10.1007/s10670-019-00212-4

Neal, Radford M (2006). Puzzles of anthropic reasoning resolved using full non-indexical conditioning. *Technical Report No. 0607*, Dept. of Statistics, University of Toronto. https://arxiv.org/abs/math/0608592

Pust, Joel (2012). Conditionalization and essentially indexical credence. *The Journal of Philosophy*, 109, 295-315. https://doi.org/10.5840/jphil2012109411

Thomas, Teruji (2021). Simulation expectation. *Global Priorities Institute*: GPI Working Paper No. 16-2021. https://globalprioritiesinstitute.org/wp-content/uploads/Teruji-Thomas-Simulation-expectation.pdf

Titelbaum, Michael G. (2012). An embarrassment for double‐halfers. *Thought: A Journal of Philosophy*, 1, 146-151. https://doi.org/10.1002/tht3.21

Weatherson, Brian (2005). Should we respond to evil with indifference? *Philosophy and Phenomenological Research*, 70, 613-635. https://doi.org/10.1111/j.1933-1592.2005.tb00417.x

1. Department of Philosophy, Dartmouth College, Hanover, NH 03755, USA [↑](#footnote-ref-1)
2. Department of Philosophy, Northeastern University, Boston, MA 02115, USA [↑](#footnote-ref-2)
3. This is not to say that my experience is *identical* whether or not I am simulated: see Neal (2006, 18) and Carlsmith (2022) for discussion of this point. [↑](#footnote-ref-3)
4. We do not mean to suggest that this kind of indifference principle is established beyond question: see Elga (2004) and Weatherson (2005) for some debate. But Bostrom and Chalmers assume it, and we take it for granted. [↑](#footnote-ref-4)
5. See Thomas (2021, p. 5). Thomas assumes that the probability that you are simulated can be equated with the frequency of simulated experiences among experiences like yours, across all worlds. Thomas’s goal is to resist Birch’s (2013) selective skepticism objection; we make no claim here about the strength of this objection. [↑](#footnote-ref-5)
6. See Lewis (2013, p. 4018). Lewis’s main concern here, though, is to compare the strength of the doomsday argument and the simulation argument. [↑](#footnote-ref-6)
7. According to Musk, the odds that you are not living in a computer simulation is “one in billions” (quoted in Chalmers 2022, p. 83). [↑](#footnote-ref-7)
8. Note that there is a distinction between Lewisian halfers (Lewis, 2001; Bradley 2011) and double-halfers (Bostrom, 2007; Meacham, 2008; Pust, 2012). Rebuttals of one position are not necessarily rebuttals of the other, as witnessed by the fact that double-halfers offer rebuttals of the Lewisian position (Meacham, 2008, p. 262), and vice versa (Bradley, 2011, p. 408). Of particular note in the present context: Bostrom (2007) offers a double-halfer solution to Sleeping Beauty, and Neal (2006, p. 14) supplies a powerful thirder rebuttal. [↑](#footnote-ref-8)
9. Elga (2004) argues for the indifference principle at the heart of his (2000) thirder approach, rather than applying thirder reasoning. However, the crucial feature of this paper for present purposes is that it makes clear that uncertainty about which person you are should be treated in the same way as uncertainty about what time it is (2004, p. 396). [↑](#footnote-ref-9)
10. Not everyone shares the view that conclusions of this kind are absurd, e.g. Hansen (2023, p. 720). [↑](#footnote-ref-10)
11. We do not wish to imply that Milano intended her approach to be taken literally. Nor should we be understood as endorsing either Milano’s (2022) or Horgan and Mahtani’s (2013) approach to conditionalization in cases involving self-location uncertainty. We merely wish to stress that updating approaches can be applied to simulation scenarios. Similar considerations apply to Dutch book and scoring rule approaches (Briggs 2010, and references therein). If one thinks of the profit/loss or the score as belonging to an individual, one might worry that these approaches do not carry over from Sleeping Beauty to cases involving distinct people. But one can think instead of the *expected* score of a person, given that the person is unsure which individual she is. [↑](#footnote-ref-11)
12. This case also has the virtue that it doesn’t involve memory loss, duplicates, or any other “fantastic assumption” that might lead us astray (Neal 2006, p. 11). [↑](#footnote-ref-12)
13. Bostrom (2007) uses this “extreme Sleeping Beauty” scenario as part of an *objection* to thirder reasoning, but he does not dispute that thirder reasoning has the mathematical consequence in this case that we rely on. [↑](#footnote-ref-13)
14. Neal doesn’t *rely* on this argument strategy, but only takes it to be “further evidence” for the thirder solution (2006, p. 15). [↑](#footnote-ref-14)
15. Recall from section 5.2 that Milano’s (2022) updating approach seemed unduly limited in that, taken literally, it presupposes a *temporal* possibility structure. This limitation is avoidable: we don’t have to interpret updating in a literal, temporal sense. However, the present limitation seems *unavoidable*: we *do* have to understand updating in terms of additional *information*, and that feature is missing in many cases in which the thirder conclusion follows. [↑](#footnote-ref-15)
16. For example, Dutch book approaches and scoring rule approaches assess the profit/loss or the score without reference to any unexperienced locations. Briggs (2010) argues for a thirder conclusion on the basis of these approaches. [↑](#footnote-ref-16)