The Doomsday Argument and the Simulation Argument

Peter J. Lewis

ABSTRACT: The Doomsday Argument and the Simulation Argument share certain structural features, and hence are often discussed together (Bostrom 2003, Aranyosi 2004, Richmond 2008, Bostrom and Kulczycki 2011). Both are cases where reflecting on one's location among a set of possibilities yields a counter-intuitive conclusion—in the first case that the end of humankind is closer than you initially thought, and in the second case that it is more likely than you initially thought that you are living in a computer simulation. Indeed, the two arguments do have some structural similarities. But there are also significant disanalogies between the two arguments, and I argue that these disanalogies mean that the Simulation Argument succeeds and the Doomsday Argument fails.

1. The Doomsday Argument

If we make some arbitrary stipulations about precisely when our ancestors counted as human, and about precisely when a person counts as being born, then every human being has a unique numerical birth rank. Presumably you don't know what your birth rank is, although you know some constraints on it: it has to be bigger than 100, for example.

Given our arbitrary stipulation about the beginning of the human race, and given some further stipulations about what would count as the end of the human race, one can also consider the question of how many human beings there will be in total. Presumably you don't know this either, although you may have some opinions on the matter. We should be able to quantify these opinions by eliciting your betting preferences. At what odds would you bet on the hypothesis that there will be between 100 and 200 billion humans in total? At what odds would you bet on there

being between 100 and 200 trillion? These odds constitute your credences in the various hypotheses about the total number of humans.

Now suppose you were to learn your precise birth rank. What effect should this have on your credences concerning the total human population? The obvious answer is "None at all". However, the Doomsday Argument suggests otherwise; it suggests that if you reason rationally, learning your birth rank should increase your credence in low total human populations at the expense of your credence in high total human populations (Leslie 1990; Bostrom 1999). And your birth rank is a simple fact about the world, so presumably your credences after learning this evidence better reflect the state of the world than your credences beforehand. Hence the name of the argument: there will be fewer humans in total than you currently think, and so doom for humanity is coming sooner than you currently think.

Let us run through the reasoning behind these assertions. To simplify matters, assume that you initially have no idea whatsoever what your birth rank is, except that you are sure it is less than some number n. That is, you are uncertain which of the n hypotheses E_1 through E_n is true, where E_1 says your birth rank is 1, and so on. And assume that initially you have no idea how many humans there will be in total, except that it is smaller than n. That is, you are uncertain which of the n hypotheses H_1 through H_n is true, where H_1 says there is 1 human in total, and so on.

Notoriously, even in this idealized case, assigning credences to these hypotheses is a controversial matter. To illustrate, consider as a toy model the case $n = 3$, shown in figure 1. There are two prima facie plausible ways you might distribute your credences over the possible self-locations shown in the grid. You might reason as follows. There are six possible locations I could occupy—I might be the only person in a one-person world, or the first person in a two-

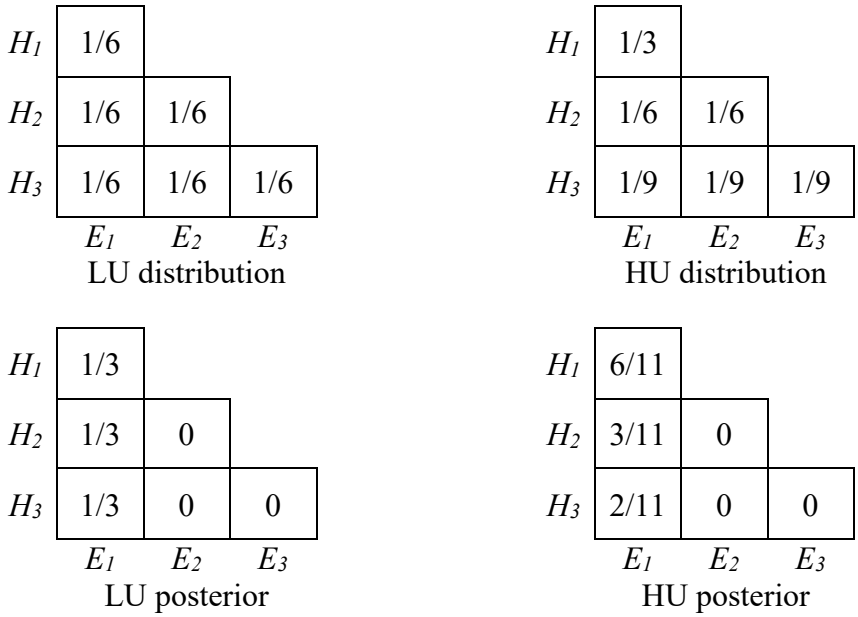

Figure 1: Doomsday Argument

person world, or the second person in a two-person world, and so on—and since I'm indifferent between them all, I should assign each a credence of 1/6.[1] These are the credences depicted in the top left of figure 1; call this the location-uniform (LU) distribution.

Alternatively, you might reason as follows. There are three possible worlds I might occupy—those described by H_1, H_2 and H_3—and since I'm indifferent between them, I should assign each a credence of 1/3. My credence in H_2 is split further between the two locations I might occupy in this world, and my credence in H_3 is split further between three such locations. Since I am indifferent between these locations, my credences in H_2 and H_3 should be split evenly between the possible locations, resulting in the credences depicted in the top right of figure 1; call this the hypothesis-uniform (HU) distribution.

Now consider what happens when I learn my birth rank. Suppose I learn that my birth rank is 1—that is, I learn that E_1 is true and E_2 and E_3 are false. How should this affect my

[1] The appeal to indifference in this paragraph and the next is for simplicity only, and plays no role in the argument. I indicate how the argument generalizes to non-uniform priors below.

credences? Straightforward Bayesian conditionalization says that the result is obtained by setting my credences in E_2 and E_3 to zero and renormalizing the remaining credences. For the LU distribution this gives the posterior credences shown in the bottom left, and for the HU distribution it gives the posterior credences shown in the bottom right.

Note that in each case, when you learn that your birth rank is 1, H_1 is confirmed: under the LU distribution, your credence in H_1 goes up from 1/6 to 1/3, and under the HU distribution, your credence in H_1 goes up from1/3 to 6/11. Similarly, your credence in H_3 goes down under either LU or HU. Learning your birth rank has made it more likely that the total human population is small, and less likely that it is large.

The above results do not depend on learning any particular birth rank; it is easy to see that if I learn that my birth rank is two, H_2 is confirmed over H_3 under either LU or HU.[2] Nor do they require that you are initially indifferent between the relevant hypotheses; the results generalize to any initial credence distribution.[3] And the argument can be further generalized to any value of n; when you learn that your birth rank is r, this evidence confirms the smaller total populations consistent with this birth rank and disconfirms the larger ones.[4] As it is, you don't know your birth rank, but whatever it is, conditionalizing on it will increase your credence in smaller total human populations over larger ones. So your current credence is too rosy; "doom soon" is more likely than you currently think.

---

[2] Under the LU distribution, your initial credences in H_2 and H_3 are 1/3 and 1/2 respectively, and your final credences are 1/2 each. Under the HU distribution, your initial credences in H_2 and H_3 are 1/3 each, and your final credences are 3/5 and 2/5 respectively. In each case H_2 is confirmed.

[3] Suppose your credences in H_1, H_2 and H_3 are p_1, p_2 and p_3, where $p_1 + p_2 + p_3 = 1$. Under HU, your initial credence in H_1 is p_1 and your final credence is p_1/q, where $q = p_1 + p_2/2 + p_3/3 < 1$, so H_1 is confirmed. Under LU, your initial credence in H_1 is p_1/q and your final credence is p_1, where $q = p_1 + 2p_2 + 3p_3 > 1$, so again H_1 is confirmed.

[4] This is easy to see qualitatively for uniform priors. Under LU, your initial credences in the H_i increase as i gets larger, but your final credences in the H_i are uniform over $i \geq r$. So your credence is redistributed to small-i hypotheses, and H_r is confirmed. Under HU, your initial credences in the H_i are uniform, but your final credences for $i \geq r$ decrease as i gets larger. So again your credence is redistributed to small-i hypotheses, and H_r is confirmed.

2. Doom and Self-Location

Is the Doomsday Argument compelling? There has been considerable debate over this (Dieks 1992, Korb and Oliver 1998, Bostrom 1999, Olum 2002, Bostrom 2002, Bostrom and Cirković 2003), but I wish to concentrate here on a response that has been largely overlooked in this literature. The criticism I wish to focus on is that, as Pisaturo (2009) has pointed out, evidence that confirms a shorter *total* duration for humanity doesn't necessarily confirm a shorter *future* duration for humanity. This can be demonstrated in our toy model by reconsidering the diagrams in figure 1. Suppose we consider hypothesis *D*, which says that you are the *last* human being (i.e. that the future duration of humanity is as short as it can be). This hypothesis is represented by the three locations along the long diagonal in each of the diagrams. So under LU, *D* initially has a credence of 1/2, but after you learn that your birth rank is 1, your credence in *D* drops to 1/3. Similarly, under HU, *D* initially has a credence of 11/18 (0.611), but after you learn that your birth rank is 1, your credence in *D* drops to 6/11 (0.545). So even though a small total duration for humanity is confirmed, a short future duration is *disconfirmed* (Lewis 2010).

This result might seem paradoxical at first, but in fact it can be readily explained. When you learn that your birth rank is 1, it becomes more likely that the total population is small, but on the other hand you also learn that you are at the beginning of whatever population you are part of. These two factors work against each other, and in the above case, the net result is that the "doom soon" hypothesis *D* is disconfirmed rather than confirmed. Call this the relative location effect.

How far can this result be generalized? It depends. Under the LU distribution, it is quite robust: it readily generalizes to any value of n, and to any birth rank up to $n/2$.[5] The opposite

[5] Under the LU distribution, each location initially has a credence of $1/(1 + 2 + \ldots + n) = 2/n(n + 1)$. Hence the diagonal hypothesis *D* initially has a credence of $2/(n + 1)$. If you learn that your birth rank is 1, *D* has a final

result holds for birth ranks greater than $n/2$: discovering your birth rank makes it more likely that D is true. This latter condition is precisely what one would expect; finding that your birth rank is in the earlier half of the range of possibilities makes it less likely that the human population will end soon, and finding that is in the later half makes it more likely. What strikes us as surprising and paradoxical about the Doomsday Argument is that *any* birth rank whatsoever is supposed to make "doom soon" more likely. This analysis shows that under LU, at least, the paradoxical result does not follow: late birth ranks (relative to what you consider the range of possibilities) make "doom soon" more likely, but early birth ranks do not.

Under the HU distribution, the relative location effect is far less robust. It holds for values of n up to 6, but for larger values of n the "doom soon" hypothesis is confirmed even by a birth rank of 1.[6] Since in realistic scenarios we are concerned with values of n in the trillions, it looks like the Doomsday Argument goes through; "doom soon" is confirmed by any birth rank whatsoever. So everything depends on the way that credences are distributed over possible self-locations: under the LU distribution the Doomsday Argument fails, but under the HU distribution it succeeds.

To see if this approach succeeds in undermining the Doomsday Argument, then, we can't avoid considering the vexed question of the proper distribution of prior probabilities in contexts of self-location uncertainty. You begin with certain credences concerning the total human population. Then when the issue of birth rank is raised, you notice that each hypothesis concerning the total human population postulates a different number of locations in the birth rank

---

credence of $1/n$, which is less than its initial credence provided $n > 1$. Hence D is disconfirmed. If you learn that your birth rank is r, D has a final credence of $1/(n - r + 1)$, which is less than its initial credence provided $n > 2r - 1$. Hence D is disconfirmed for any birth rank less than $(n + 1)/2$.

[6] Under the HU distribution, the locations in row i initially have credences of $1/in$. Hence the diagonal hypothesis D has an initial credence of $(1 + 1/2 + \ldots + 1/n)/n$, and if you learn that your birth rank is 1, D has a final credence of $1/(1 + 1/2 + \ldots + 1/n)$. Numerical solution shows that D is disconfirmed for $1 < n \leq 6$, and confirmed for $n \geq 7$.

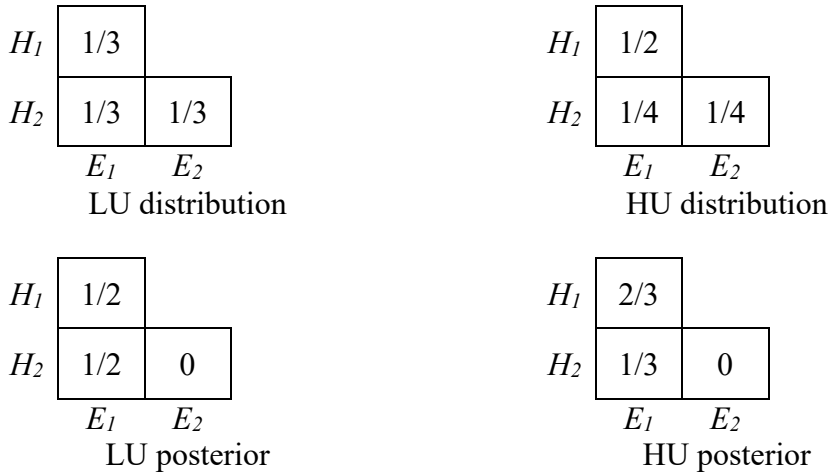

Figure 2: Sleeping Beauty (or $n = 2$ Doomsday Argument)

sequence you might occupy. How should you distribute your credence over these possible self-locations?

This question has been extensively investigated in the context of the Sleeping Beauty puzzle (Elga 2000), which has a very similar structure to the Doomsday Argument. Indeed, for $n = 2$ the structure of possibilities in the Doomsday case is identical to that in the Sleeping Beauty case—the structure shown in figure 2. In the Sleeping Beauty case, H_1 and H_2 are the hypotheses that a fair coin shows heads and tails respectively, and E_1 and E_2 are the hypotheses that it is Monday and Tuesday respectively. Beauty is initially uncertain whether it is Monday or Tuesday, but later learns what day it is. The puzzle concerns the credence Sleeping Beauty should ascribe to heads before she learns what day it is. The LU distribution corresponds to the answer 1/3—the "thirder" solution to the puzzle—and the HU distribution to the answer 1/2—the "halfer" solution.

The halfer solution is unpopular, and with good reason. The main count against it is that Sleeping Beauty *ends up* with a credence of 2/3 in heads, after she learns what day it is. Since one can suppose that the coin is tossed on Monday night, this amounts to Sleeping Beauty

ascribing a credence of 2/3 to heads for a future toss of a coin she acknowledges to be fair. This looks prima facie irrational. Admittedly, under the thirder solution, before she is told what day it is Sleeping Beauty has a credence of 1/3 in heads for a future toss of a fair coin. But before she is told what day it is, Sleeping Beauty is subject to a peculiar kind of self-location loss, where her possible locations are distributed unevenly between heads and tails. That is, before she is told what day it is, she has an epistemic excuse for her peculiar credence in heads, but afterwards she has no such excuse.

Exactly the same argument can be used against the HU distribution. In the $n = 2$ case, after you learn your birth rank, your credence that the total human population consists of a single person (H_1) is 1/2 under the LU distribution but 2/3 under the HU distribution. Suppose that the gods wait until the end of the first person's lifespan to decide whether or not to create a second person, and they make this decision using a fair coin, heads for "no" and tails for "yes". Then if you adopt the HU distribution your credence in heads for a future toss of a fair coin is 2/3, which is prima facie irrational. Of course, the number of people may not actually be determined by gods rolling dice, but it could be determined this way, and the HU distribution plainly gives the wrong posterior probability.

Under the LU distribution, your posterior credence in heads is 1/2, as it should be. But your prior credence is 1/3, and this is still a credence in heads for a future toss of a fair coin. However, as in the Sleeping Beauty case, you have an excuse for your anomalous credence: you don't know which person you are, and your possible locations are distributed unevenly between heads and tails. After you learn your birth rank (and hence reduce your possible locations so that they are evenly distributed between heads and tails) you have no excuse for any deviation from 1/2. And there is nothing special about the $n = 2$ case: the same reasoning applies for any value

of n. The HU distribution leads to a posterior credence in H_1 that is greater than its objective chance (given that the gods choose a total population at random), whereas the LU distribution does not. Hence the LU distribution is correct and the HU distribution is incorrect.

In the Sleeping Beauty case, the thirder solution amounts to the position that your credence in heads after you learn what day it is should be the same as your credence in heads on Sunday, before the self-location loss occurs. In addition to being a consequence of the above argument, this position is independently plausible; surely gaining and then losing some self-location uncertainty should leave your credences unchanged. Similarly in the Doomsday case, under the LU distribution your posterior credences in the H_i are the same as your credences before you took into account your own location in the population. Under the HU distribution, your posterior credences differ from those you started with. Again, it is independently plausible that introducing and then eliminating some uncertainty regarding your own location should leave your credences unchanged, as the LU distribution entails.

If this argument for the LU distribution is correct, then there are two senses in which the Doomsday Argument fails, depending on what one takes its conclusion to be. If one takes the conclusion of the Doomsday Argument to be that any birth rank whatsoever confirms a short *future* duration for humanity, then the relative location effect shows that this conclusion does not in fact follow. Some birth ranks confirm a short future, but others disconfirm it, exactly as one would expect. Alternatively, if one takes the conclusion of the Doomsday Argument to be that any birth rank whatsoever confirms a small *total* duration for humanity, then the above argument shows that it fails. If learning your birth rank does not restore the prior probabilities concerning the total human population that you started with, then you distributed your credences in an irrational way.

It might be objected that I beg the question against the Doomsday Argument here: I assume that gaining and losing self-location uncertainty cannot shift one's credences in the H_i, and yet that very shift is the conclusion of the Doomsday Argument. A couple of responses are in order. First, it depends on what one takes the conclusion of the Doomsday Argument to be. If one takes the Doomsday Argument to be about the future duration of humanity, then I do not assume that the conclusion of the argument is false; rather, I *show* that it is false, based on an assumption about your credences concerning the total duration of humanity. Second, even if one takes the Doomsday Argument to be about the total duration of humanity, I do not *assume* that your initial and final credences must be the same, but rather argue for this conclusion based on structural similarities with the Sleeping Beauty puzzle.[7]

If this is right, only a direct argument against the rationality of the LU distribution could serve to defend the Doomsday Argument. And indeed such arguments have been given, most notably Bostrom's "presumptuous philosopher" argument (2002, 124). Bostrom asks us to consider a situation in which physicists have determined that there are two possible cosmological theories, according to one of which there are a trillion trillion observers, and according to the other of which there are a trillion trillion trillion observers. The physicists propose an expensive experiment to determine which is correct, but the presumptuous philosopher informs them that no such experiment is necessary; a simple appeal to LU entails that the latter is a trillion times more credible than the former. This is intended as a reductio of the LU distribution.

But this example as it stands doesn't work: it doesn't introduce any self-location uncertainty in addition to the uncertainty about which cosmological theory is true. The presumptuous philosopher is not uncertain about which observer she is; she can point to any

[7] Admittedly, even though the thirder solution to the Sleeping Beauty puzzle is widely held, it can be challenged. So a more circumspect conclusion so far would be that *if* the thirder solution is right, then the Doomsday Argument fails (Dieks 1992). But this is still interesting; compare footnote 11 below.

number of properties and relations that uniquely identify herself among all the conscious beings that exist. The same goes, presumably, for (almost?) all the observers in these possible universes; they are not uncertain which observer they are, so there is no call for distributing their credences over possible observers, in the LU manner or otherwise.[8]

One could imagine a variant of this argument in which there is genuine self-location uncertainty; indeed, the choice between standard and many-worlds quantum mechanics seems to afford such a case. In this case, many-worlds quantum mechanics says that observers are constantly branching into a huge number of subjectively identical copies, whereas standard quantum mechanics says that a single observer remains a single observer. Here an observer acquires genuine self-location uncertainty under one hypothesis but not the other, and a presumptuous philosopher may well argue that the many-worlds hypothesis is automatically more credible because of this uncertainty under LU (Price 2008, 12). But this is not clearly a reductio of LU, since it remains contested whether the very notions of probability and decision under uncertainty make any sense for the many-worlds theory.[9] The presumptuous philosopher argument might well be a reductio of the many-worlds hypothesis rather than LU (Price 2008, 14).

I conclude that the Doomsday Argument is far from compelling. So let us compare it to the Simulation Argument and see to what extent the problems that afflict the former also afflict the latter.

[8] Bostrom's actual target is the Self-Indication Assumption, which says "Given the fact that you exist, you should (other things being equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist" (Bostrom 2002, 66). This assumption makes no mention of self-location uncertainty, and the presumptuous philosopher argument may well be telling against it. But this just shows that the SIA is far too general; taken as an objection to LU, the presumptuous philosopher argument is ineffective.
[9] See the papers in sections 3 and 4 of Saunders et al. (2010) for arguments on both sides of this issue.

3. The Simulation Argument

It is possible that technology could advance to a point such that the lives of conscious being like ourselves could be simulated on a computer. Indeed, it is possible that technology has reached that point already, and you are living in such a simulation. What credence should you attach to this possibility? Presumably you ascribe it an extremely low credence; the import of the Simulation Argument is that this credence should be revised upwards.[10]

There are a number of forms the Simulation Argument can take, depending on the precise structure of the simulation. Let us start by considering the structure than makes the Simulation Argument most closely analogous to the Doomsday Argument. Let H_1 be the hypothesis that there are no simulated worlds, just the physical world. Let H_2 be the hypothesis that the physical world contains exactly one simulated world, where the two worlds contain roughly the same number of conscious beings. Let H_3 be the hypothesis that the physical world contains exactly one simulated world, and that simulated world contains its own simulated world, where all three worlds—the physical world, the first-level simulation and the second-level simulation—contain roughly the same numbers of conscious beings. Suppose that these are all the possibilities. Finally, let E_1, E_2 and E_3 be the self-location hypothesis that you live in the physical world, the first-level simulation and the second-level simulation respectively.

[10] Bostrom (2003) in fact argues for a disjunctive thesis: either the human species is very likely to go extinct before developing the required technology, or any civilization with such technology is extremely unlikely to run a significant number of simulations, or we are almost certainly living in a simulation. What Bostrom calls "the core of the simulation argument" is the argument that if the first two disjuncts are false, then you should revise your credence that you are living in a simulation upwards to almost 1. It is this core argument that I address here.

How should you distribute your credences over these possibilities? First of all, it depends what your initial credences in the hypotheses H_i are. Let us suppose—implausibly, but we will relax this assumption later—that you initially ascribe equal credence to the three H_i. As in the Doomsday Argument, each hypothesis H_i entails a different number of locations you might occupy, as shown in figure 3. Again, you might redistribute your credences evenly over these locations—the LU distribution—or you might just subdivide them evenly within each of the H_i—the HU distribution. These credences are shown in the diagrams at the top of figure 3.

Note that the LU distribution, which I have been advocating, has the effect of shifting your credence away from H_1 and towards H_3. That is, applying LU seems to involve something like a disconfirmation of the hypothesis that there are no simulated worlds, thereby making it more likely than you initially thought that you live in a simulated world.

But such a conclusion would be too hasty. As Richmond (2008, 212) has noted, you have evidence that eliminates your self-location uncertainty, because you know that our world does not contain any simulated worlds; we don't have the technology to simulate the lives of

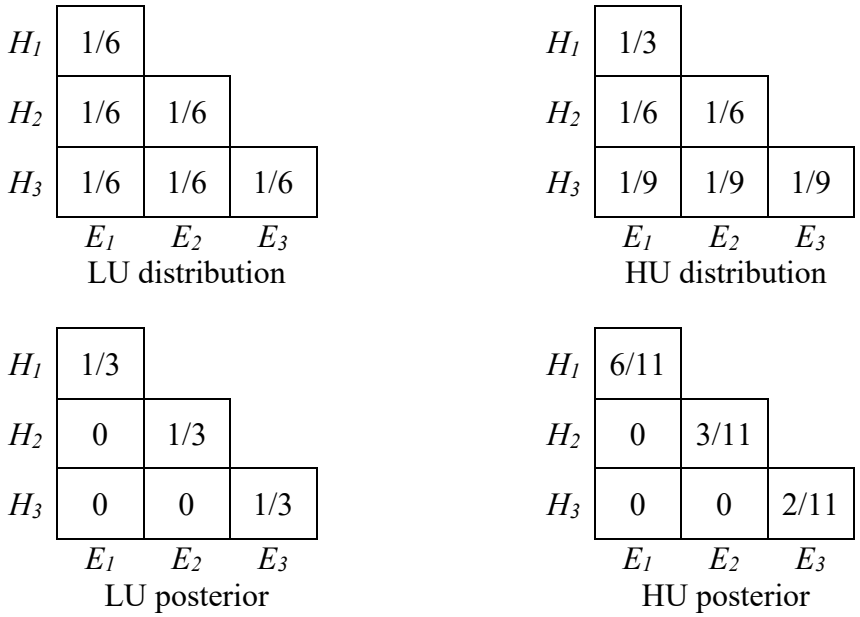

Figure 3: Simulation Argument

conscious beings. If the possibilities have the structure shown in figure 3, this evidence has the effect of ruling out any locations other than those that lie on the main diagonal. As in the Doomsday case, we can apply straightforward Bayesian conditionalization to see what this evidence should do to your credences: set all the off-diagonal elements to zero, and renormalize. This results in the credences shown in the bottom row of figure 3. The effect (for both the LU and HU distributions) is to confirm H_1. This is the move that is analogous to the Doomsday Argument: just as the Doomsday Argument suggests that a small human population is more likely than you thought, so this argument suggests that *absence* of simulation is more likely than you thought. Since it is less likely than you thought that there are simulations, one might conclude that it is also less likely than you thought that you are living in a simulation.

So a straightforward analogy between the Doomsday Argument and the Simulation Argument might suggest that the Simulation Argument works in the opposite of the advertised direction; learning your location in the simulation hierarchy makes it *less* likely that you are living in a simulation, not more likely. But this conclusion too would be hasty, for two distinct reasons.

First, as in the Doomsday Argument, the effect of your evidence is to restore your credences in the H_i to their initial values, at least under LU. And LU seems just as compelling in this case as in the Doomsday case. You notice that the hypotheses H_i introduce uncertainty concerning your own location in the hierarchy, and then notice that the evidence that this world contains no simulations takes away this self-location uncertainty (and tells you nothing else). Surely this should restore your prior credences in the H_i, as LU (but not HU) entails. And as above, one could appeal to an analogy with the Sleeping Beauty puzzle to provide further support for the LU distribution. The analogy is not so close in case of the Simulation Argument;

the gods cannot decide after the fact whether you are a simulation or not. But the gods can still decide whether to produce a simulation based on a coin-toss, and the point of the Sleeping Beauty argument is that it makes no difference whether the coin toss is in the future or the past: when fully informed about your location, your credence in heads should be 1/2, and hence the LU distribution must be correct. But in that case, the net effect of self-location considerations on your credences in the H_i is necessarily zero. The Simulation Argument does not have the opposite effect from the one advertised, because it has no effect at all.

The second reason that the above conclusion is too hasty is the relative location effect. Recall Pisaturo's point concerning the Doomsday Argument—that evidence for a shorter total duration for humanity is not necessarily evidence for a short future duration. Exactly the same point can be made here: evidence against the existence of simulations is not necessarily evidence that you are not living in a simulation. In fact, it is easy to show that in the case we are considering, the evidence is simultaneously evidence for absence of simulation *and* evidence that you are living in a simulation.

To see this, consider what happens to your credences when you update them on the evidence that this world does not contain a simulation. In particular, consider the effect on the hypothesis E_1 that you live in a physical world rather than a simulated one. Under the LU distribution your credence in this hypothesis goes from 1/2 to 1/3, and under the HU distribution it goes from 11/18 (0.611) to 6/11 (0.545). In each case, the evidence that this world does not contain a simulation makes it less likely that you live in a physical world, and more likely that you live in a simulated one.

As in the Doomsday case, this might be considered paradoxical, but it can be readily explained. The evidence that this world contains no simulations makes it more likely that there

are no simulations, but *if* there are simulations, it makes it certain that you live in one. These two factors act against each other, and in the $n = 3$ case, the result is that the evidence makes it less likely that you live in a simulation, under either an LU or an HU distribution.

How robust is this effect? Given the exact symmetry between this case and the Doomsday case (simply exchange hypotheses E_1 and D), it follows that the results are the same as before. That is, under the LU distribution the effect generalizes to any value of n, and under the HU distribution, the effect holds only for values of n up to 6. Since I take the LU distribution to be the correct one for the reasons given above, I take the effect to be robust.[11]

So what does all this mean for the Simulation Argument? As in the Doomsday case, it depends what you take the conclusion of the argument to be. If you take the conclusion to be that self-location considerations disconfirm H_1 (the hypothesis that there are no simulations), then the Simulation Argument fails. The initial disconfirmation of H_1 that occurs when you distribute your credences over the possible self-locations in the simulation hierarchy is exactly counteracted by the shift when you conditionalize on the evidence that this world contains no simulations. The net effect is zero.

However, if you take the conclusion of the Simulation Argument to be that conditionalizing on evidence concerning your self-location disconfirms E_1 (the hypothesis that you do not live in a simulation), then the argument succeeds. Thanks to the relative location effect, conditionalizing on the fact that the world contains no simulations makes it more likely that you live in a simulation. Recall that the relative location effect undermines the Doomsday Argument. The reason for the difference is that the conclusions of the two arguments target

[11] A more circumspect conclusion is that *if* the thirder position is correct, then the Simulation Argument succeeds. But recall that the equivalent conclusion for the Doomsday Argument is that it *fails* if the thirder position is correct. Even in this conditional form, the conclusion is interesting: the two arguments should not be taken as simply two instances of the same form of reasoning.

complementary regions of the space of possible self-locations; the "doom soon" conclusion claims that you inhabit the extreme fringe of the space of possibilities (i.e. the main diagonal), whereas the "simulation" conclusion claims that you *do not* inhabit the extreme fringe (you are not in the left column). Hence despite the superficial similarity of the two arguments, there is an important sense in which they are opposed to each other: one succeeds when the other fails.

So the Simulation Argument seems to be on firmer ground than the Doomsday Argument. But a second disanalogy between the two arguments might be taken to undercut this advantage. In the Doomsday case, it is plausible that you do not initially know your birth rank, and hence it makes sense to conditionalize on this information. In the Simulation case, though, it is plausible that you already know that this world contains no simulations, so it is not appropriate to conditionalize on the information. In that case, the relative location effect does not arise. Instead, the bottom-left diagram in figure 3 represents your prior credences, and since you learn nothing, these are your final credences too.

Still, even if the typical agent already knows that the world contains no simulations, it is instructive that if one didn't know this, learning it could confirm the hypothesis that one is living in a simulation. Furthermore, we have been considering only the version of the Simulation Argument that is most closely analogous to the Doomsday Argument; considering less analogous versions strengthens the claim that the Simulation Argument succeeds where the Doomsday Argument fails.

4. Alternative Simulation Structures

To that end, let us consider an alternative version of the Simulation Argument. In the version we have been considering up to now, each world, if it contains a simulation, contains exactly one

simulation with roughly the same number of inhabitants as the simulating world. And we have been assuming that simulated worlds can themselves contain simulated worlds. But suppose instead that a world can contain several simulations (where, for simplicity, we still assume that each simulation contains roughly the same number of inhabitants as the simulating world). And suppose that simulated worlds cannot themselves contain simulated worlds; there is at most a two-level hierarchy of simulations.

Then we can consider the following hypotheses. H_1 is the hypothesis that there are no simulated worlds, just the physical world. H_2 is the hypothesis that the physical world contains one simulated world, where the two worlds contain roughly the same number of conscious beings. H_3 is the hypothesis that the physical world contains two simulated worlds, where all three worlds roughly the same number of conscious beings. It is easy to see how this sequence could be continued to any number of hypotheses, but for simplicity let us stick to the $n = 3$ case, and for further simplicity let us suppose that you are initially indifferent between H_1, H_2 and H_3. Then your credences, under the LU and HU distributions, are exactly the same as before; they are shown in the top row of figure 4.

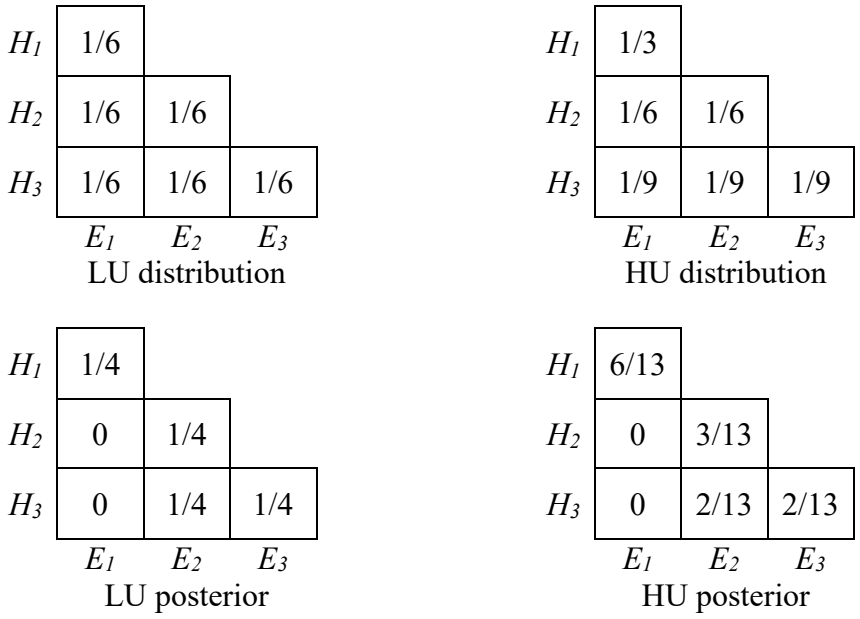

Figure 4: Simulation Argument (two-level)

The difference is what happens when you conditionalize on the evidence that our world contains no simulations. In this case, rather than restricting your location to the main diagonal, this evidence just rules out the first column below the first row. Conditionalizing yields the credences in the bottom row of figure 4. Again H_1 is confirmed under either prior; the evidence confirms the hypothesis that there are no simulations. Furthermore, the hypothesis that you are living in a simulation is confirmed under either prior. Under the HU distribution your credence that you are living in a simulation (i.e. not E_1) goes from 7/18 to 7/13, and under the LU distribution it goes from 1/2 to 3/4. This effect is stronger than in the original version.

Significantly, in this version of the Simulation Argument your self-location uncertainty is never completely eliminated. If H_3 is true, learning that our world contains no simulations doesn't discriminate between location E_2 (the first simulated world) and E_3 (the second simulated world). This residual self-location uncertainty means that the shift away from H_1 that occurs when you apply the LU distribution is not completely undone when you learn that this world contains no simulations. Or alternatively, if you already know that the world contains no simulations, the application of LU to your prior credences leads directly to those shown in the bottom-left diagram of figure 4. Either way, the hypothesis that you are living in a simulation is more likely than you thought.

How strong is this effect? It depends on your prior probabilities in the H_i and your estimate of the value of n (the maximum possible number of simulated worlds that a physical world can support). If n is large, the effects are striking. Suppose, for example, you think that it is possible that the physical world could support up to a million simulated worlds, i.e. $n = 10^6$. And suppose that initially you are 99% sure that H_1 is true, with your remaining credence divided evenly over the remaining hypotheses H_2 through H_n. Then when you distribute your

credence over your possible self-locations according to LU, your credence in H_1 drops from 99% to 0.02%.[12] Conditionalizing on the evidence that this world contains no simulations has negligible effect on your credences.[13] So whether you regard this as new evidence or not, the net effect of self-location considerations on your credences is to drastically decrease you confidence that there are no simulations, resulting in a final credence that you are living in a simulation of 99.98%.

The Simulation Argument can be ramified in various ways: most obviously, one can combine the assumption that each world can contain multiple simulations with the assumption that a simulated world can contain a higher-order simulation. But such complications would only serve to reinforce the point just made via this simple calculation: the power of the Simulation Argument is that it involves self-location uncertainty that is not resolved by the evidence at hand. Hence there are two crucial disanalogies between the Simulation Argument and the Doomsday Argument. First, the relative location effect supports the Simulation Argument but undermines the Doomsday Argument. Second, the self-location uncertainty in the latter can be (and by hypothesis is) fully resolved, so there should be no net shift in our credences. Together, these differences make the Simulation Argument far more convincing than the Doomsday Argument.

References

Aranyosi, István A. (2004) "The Doomsday Simulation Argument. Or why isn't the end nigh, and you're not living in a simulation", http://philsci-archive.pitt.edu/1590/.

[12] Since the prior probabilities are not uniform, we need to use a generalized LU distribution. That is, if your prior probabilities in the hypotheses H_i are p_i, your credence in each possible self-location along the H_i row is ap_i, where a is a constant given by $\Sigma_i iap_i = 1$. In this case, p_1 is 0.99, p_2 through p_n are 10^{-8}, and n is 10^6, resulting in a value for a of 1/5001. Hence your credence in H_1 becomes $ap_1 = 0.02\%$.

[13] The locations eliminated by this evidence (the left-hand column below the top row) have a total credence of $a(p_2 + p_3 + \ldots + p_n)$, which is of the order of 10^{-6}. Hence a negligible proportion of your total credence is redistributed by this evidence.

Bostrom, Nick (1999), "The doomsday argument is alive and kicking", *Mind* 108: 539–551.

——— (2002), *Anthropic Bias: Observer Selection Effects in Science and Philosophy*. New York: Routledge.

——— (2003), "Are you living in a computer simulation?", *Philosophical Quarterly* 53: 243–255.

Bostrom, Nick and Milan M. Cirković (2003), "The Doomsday Argument and the self–indication assumption: reply to Olum", *Philosophical Quarterly* 53: 83–91.

Bostrom, Nick and Marcin Kulczycki (2011), "A patch for the Simulation Argument", *Analysis* 71: 54–61.

Dieks, Dennis (1992), "Doomsday—or: The dangers of statistics", *Philosophical Quarterly* 42: 78–84.

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

Korb, K. B. and J. J. Oliver (1998), "A refutation of the doomsday argument", *Mind* 107: 403–410.

Leslie, John (1990), "Is the end of the world nigh?", *Philosophical Quarterly* 40: 65–72.

Lewis, Peter J. (2010), "A note on the Doomsday Argument", *Analysis* 70: 27–30.

Olum, Ken D. (2002), "The Doomsday Argument and the number of possible observers", *Philosophical Quarterly* 52: 164–184.

Pisaturo, Ronald (2009), "Past longevity as evidence for the future", *Philosophy of Science* 76: 73–100.

Price, Huw (2008), "Probability in the Everett world: Comments on Wallace and Greaves", http://philsci-archive.pitt.edu/2719/.

Richmond, Alasdair M. (2008), “Doomsday, Bishop Ussher and simulated worlds”, *Ratio* 21: 201–217.

Saunders, Simon, Jonathan Barrett, Adrian Kent and David Wallace (2010) (eds.), *Many Worlds: Everett, Quantum Theory and Reality*. Oxford: Oxford University Press.