Beyond the Doomsday Argument: Reply to Sowers and Further Remarks

NICK BOSTROM

George Sowers tries to refute the Doomsday argument on grounds that true random sampling requires all possible samples to be equally probable the time when the sample is taken. Yet the Doomsday argument does not rely on true random sampling. It presupposes random sampling only in a metaphorical sense. After arguing that Sowers' critique fails, I outline my own view on the matter, which is that the Doomsday argument is inconclusive and that by developing a theory of observation selection effects one can show why that is so.

1. Introduction

Like Middle East conflict, the Doomsday argument has proven difficult to resolve. But there are still optimists, myself included, who hope that the issue will one day be settled.

The latest initiative comes from George Sowers Jr. In a recent paper in this journal (Sowers Jr. 2002), he claims that the Doomsday argument fails because it overlooks the correlation that exists between the cumulative human population figure at a given time and the birth-rank of the person doing the reasoning at that time. The Doomsday argument rests on the idea that one should reason as if one were a random sample from the class of all observers who will ever have existed, or at least from some considerable subset thereof that is not restricted to observers living at the current time. Yet, writes Sowers, "the strictures of random sampling … require that all results are equally probable at any time a sample is taken" (p. 40).

To this challenge, a supporter of the Doomsday argument can reply that the argument does not presuppose that you are *literally* a random sample, in an objective sense. The Doomsday argument does not allege that there is some kind of physical randomization mechanism (a time-traveling quantum stork?) that stochastically distributes preexisting observers across spacetime. Had it said something that silly, it would surely not have stirred up so much debate.

Rather, what drives the Doomsday argument is the Self-Sampling Assumption, which says you should in certain respects reason *as if* you were a random sample from all observers (or some suitable subset of observers, the observers in your "reference class"). What this amounts to is simply a prescription for your prior credence function. In the easiest case, the Self-Sampling Assumption prescribes that you should assign a prior credence of p to the hypothesis that you are an observer with property P, given that the fraction of all observers in your reference class who have property P equals p:

$$Cr(\text{I have } P \mid \text{A fraction } p \text{ of all observers in my reference class have } P) = p$$

Since the Doomsday argument does not presuppose, and the Self-Sampling Assumption does not assert, that you are an objectively random sample, a critique cannot

rest content by pointing out that the Doomsday situation doesn't satisfy "the strictures of random sampling". The reason why the Doomsday argument is controversial is precisely because it purports to show that you should apply the credence assignment given by the Self-Sampling Assumption even though you are not an objectively random sample. To determine whether the Doomsday argument is sound, we must therefore examine what grounds can be provided for thinking that the Self-Sampling Assumption applies to the doomsday situation. If the prior credence assignments it recommends are, on reflection, acceptable, then assumptions about random sampling have no further role to play and strictures of objective randomness are beside the point. If, on the other hand, the Self-Sampling Assumption's credence assignments are rejected, the Doomsday argument stops dead in its track.

One useful method of investigating the plausibility of the required prior credence assignments, which is also invoked frequently in the relevant literature, is by means of thought experiments. Following this tradition, Sowers critique comes mounted on a novel gedanken.

2. Counting balls from an urn

Suppose that two urns have been placed in front of you, one containing ten balls and the other one million balls.

> [You] have been asked by your boss to determine the number of balls in one of the urns. You decide that the best approach is simply counting the balls. You begin taking balls out one by one, setting them aside, counting as you go. After a minute or two, your boss returns and asks, 'What is your answer?' 'Does the urn contain 10 or 1,000,000 balls?' At this point you have only counted seven balls. What can you say? Simply that the urn contains at least seven balls. Analogously, for doomsday you can say only that at least 60 something billion people will ever have been born. That statement is entailed by both *FEW* and *MANY* so long as N_{FEW} is greater than 60-something billion. Hence the likelihood is one in both cases, no probability shift can occur and the conclusions of DA are avoided. (p. 41)

Here, *FEW* and *MANY* are two hypotheses about the total number of humans who will ever have existed, and N_{FEW} is the number of humans who will have existed if the more pessimistic of these hypotheses is true. The ball-counting situation is meant to be analogous to the Doomsday argument.

In the Doomsday argument, we are urged to see some support for *FEW* in the fact that our own birth ranks are about 60 billion. If we assume that the Self-Sampling Assumption is applicable to the doomsday situation, then the conditional probability of having a birth rank of 60 billion is greater given *FEW* than given *MANY*, since a greater fraction of all people that will ever have lived have birth rank 60 billion if *FEW* is true than if *MANY* is true ($1/N_{FEW}$ as opposed to $1/N_{MANY}$). From this it then follows via Bayes' theorem that *FEW* will gain in posterior probability compared to *MANY* when we take out birth ranks into account. (It does *not* follow, of course, that *FEW* ends up being

more probable than *MANY*, since that depends also on the prior probabilities of *FEW* and *MANY*. The Doomsday argument argues that there arises some *probability shift* in favor of *FEW* from taking properly into account the full evidential import of the finding that our birth ranks are lower than N_{FEW}. The Doomsday argument, thus, does not say that doomsday is likely to strike soon, only that the risk that it will strike soon has been systematically underestimated because hitherto we failed to realize that it would be more likely that we should find ourselves with birth ranks of 60 billion if a total of, say, 200 billion human will have existed than if there will be total of, say, 200 trillion humans; in the latter case our current birth ranks would be highly exceptional, which, by the Self-Sampling Assumption, is something we should reckon improbable.)

Let us examine the urn case more closely to see if it is analogous to the doomsday case by considering the sampling density for n_{count}, the random variable representing the number of balls that have been withdrawn from the urn at the time of the boss's return. The sampling density for this variable, $P(n_{count})$, depends on the sampling density $P(t_{boss})$ of the variable representing the time of the boss' return, t_{boss}.

Under natural conditions, I believe that $n_{count}=10$ would get an especially high prior probability. Why? Because if the urn contains only ten balls, which has a prior probability of one half, then $n_{count}=10$ will automatically result if the boss fails to return while the first nine balls are being counted. Therefore, unless the boss was bound to return very soon, the outcome $n_{count}=10$ will get disproportionately high sampling density.

On the other hand $n_{count}=1{,}000{,}000$ has almost no chance of occurring, unless there is a significant probability that the boss will be away for a *very* long time. And if *that* is the case, then $n_{count}=1{,}000{,}000$ presumably has a vastly greater chance of occurring than, say, $n_{count}=999{,}999$. For surely, the boss would then be more likely to return at some point in the open-ended interval that starts when you've finished counting rather than in the brief interval between withdrawal of ball number 999,999 and withdrawal of the last ball. This has the peculiar effect that the urn example, far from helping Sowers' aim of allaying our worries about being quite near the end, seems rather to suggest a starkly ultra-violent doomsday prediction: that you should rationally assign a disproportionately large prior probability to you being *the very last* human ever to be born!

Clearly this is not the intended lesson, and of course the thought experiment does not succeed in warranting the conclusion that you are strikingly likely to be the last person. Yet if we remedy the unintended anomaly that some values of n_{count} get disproportionate sample weight, for example by stipulating that any value of n_{count}, up to the total number of balls in the urn, is equally likely to be the outcome,

$$P(n_{count}=i)=\begin{cases} 1/10 & \text{if } i\le 10 \text{ and } N_{urn}=10 \\ 1/1{,}000{,}000 & \text{if } i\le 1{,}000{,}000 \text{ and } N_{urn}=1{,}000{,}000 \\ 0 & \text{otherwise} \end{cases}$$

then it easy to show that a probability shift in favor of $N_{urn}=10$ occurs upon finding that n_{count} is between 1 and 10 – a probability shift precisely mirroring the shift in favor of "few humans" that the Doomsday argument seeks to persuade us to make.

And one could argue that, in order for the urn example to be truly analogous to the Doomsday situation, one should stipulate that $P(n_{count}=i)$ be a uniform distribution in the interval [1, N_{urn}]. For in the doomsday situation, it is impossible to find that one's birth rank n_{birth} exceeds N_{total}, the total number of people that will ever have existed, and, moreover, there is no ground for assigning a greater *a priori* credence to having some particular birth rank, within the interval [1, N_{total}], than to having some other rank within that interval. This would be in accordance with applying the Self-Sampling Assumption to the doomsday situation and taking into account the observation selection effect that precludes observing values of n_{birth} that are greater than N_{total}. (In fact, as we shall see in the next section, an even weaker assumption would suffice for generating a doomsday-like probability shift.)

Therefore, while in the counting balls from an urn thought experiment it is true that we do not get a probability shift in favor of *FEW* upon finding that we have counted fewer than N_{FEW} balls when the counting stops, this observation does not refute the Doomsday argument because of the important disanalogies that exist between the two cases. In the urn case, as we saw, there is an especially great prior probability of sampling the number corresponding to the last ball counted being the very last ball in the urn. Yet surely no such bias is present in the doomsday situation; you are not a priori (before knowing what your birth rank is) especially likely to find yourself as the very last of all persons who will ever have lived.

Another defect of the urn analogy is that it does not capture the observation selection effects that operate on the corresponding evidence in the doomsday situation. Nobody observes a time prior the birth of the first observer or subsequent to the death of the last observer. In the urn case, times both before the beginning of the counting and after its conclusion are observed.

Furthermore, the "special sample", n_{count}, is the same for everybody in the urn case whereas in the doomsday situation the sample value, n_{birth}, is a different one for each observer, since we all have different birth ranks.

The urn analogy is therefore flawed in many ways.

3. The Amnesia chamber

In order to ferret out what prior probabilities one should assign to one's having a particular birth rank, given different hypotheses about the total number of humans, we may do better by turning to some other thought experiments that are more closely analogous to the doomsday situation. One such gedanken, which Sowers discusses, is *Amnesia chamber*, first introduced in (Bostrom 1997).[1] In the amnesia chamber you are supposed to be in a state of ignorance as to your birth rank. Perhaps you have been given a drug or hypnotized so as to make you forget either what year it is now, or how long it was since our species arose, or what past population figures were like. You are, however, credibly informed that there are two mutually exclusive possibilities for the human race: either the total number of humans that will ever have lived is quite small, say, 100 billion

[1] Sowers suggests that Amnesia chamber was deigned as a counterargument against certain objections against Richard Gott's version of the Doomsday argument, but that is not the case. (I have criticized Gott's version, and more generally his "delta-*t* method" elsewhere, e.g. (Bostrom 2002).)

(*FEW*), or else the total is large, 100 trillion (*MANY*). Based on whatever information you have retained in your state of partial amnesia, you assign these two hypotheses the prior credences *Cr*(*FEW*) and *Cr*(*MANY*), respectively.[2]

Next, suppose you obtain a new piece of evidence. You learn that $n_{birth} > N_{FEW}$. Of course, this conclusively proves *MANY* and raises the probability of *MANY* to unity. But note that from the fact that $n_{birth} > N_{FEW}$ raises the probability of *MANY* it follows that if we have instead conditionalized on $n_{birth} \leq N_{FEW}$ this would have increased the probability of *FEW*. For it is easily shown that you would be incoherent if you thought that conditionalizing on a new finding could lower but never raise the probability of *FEW*. (Intuitively, this is because if you thought that you were about to get a piece of evidence that could make *FEW* more probable but couldn't make it less probable, then you ought already to revise *Cr*(*FEW*) downwards, even before the new evidence comes in.) We can see this by calculating the ratios of the posterior credences of *MANY* and *FEW*. If we let *e* be the statement that $n_{birth} \leq N_{FEW}$, we have

$$\begin{aligned}\frac{Cr(MANY \mid e)}{Cr(FEW \mid e)} &= \frac{Cr(e \mid MANY)Cr(MANY)}{Cr(e \mid FEW)Cr(FEW)} \\ &= \frac{Cr(e \mid MANY)Cr(MANY)}{Cr(FEW)} \\ &< \frac{Cr(MANY)}{Cr(FEW)}.\end{aligned}$$

This follows from Bayes' theorem together with the facts that $Cr(e \mid FEW) = 1$ and $Cr(e \mid MANY) < 1$.

It is true that in the actual case, by contrast to in the amnesia chamber, there may never be a time at which we are ignorant about our birth rank while we are contemplating what credence to assign to different hypotheses about our species' prospects. But it is not clear that this discrepancy would warrant treating the cases differently. For after learning your birth rank in the amnesia chamber, your total evidence is equivalent to the evidence you have in your actual situation. Since it is hard to see, in a case like this, how the order in which you received the information could be relevant to what credence assignments you should make when all the information is in, it seems plausible that your actual credence assignments should agree with those you now think you should make upon exiting the amnesia chamber. The Amnesia chamber is merely a heuristic for considering your evidence one piece at a time, which may reduce the risk of missing some of the probabilistic implication that are be embedded in the evidence.

If the Amnesia chamber is viewed as relevantly similar to the doomsday situation, and if you are committed to updating your beliefs by Bayesian conditionalization, it would consequently *appear as if* a probability-shift, as expressed by the above inequality, should be undertaken in favor of hypotheses that indicate that doomsday will happen

[2] We could always tell a story that would justify such a credence assignment. Perhaps you have been informed that a huge fusion bomb has been buried in the center of the Earth since the beginning and that, based on some random mechanism, it is set to go off and kill the entire human species after either 100 billion or 100 trillion humans have been born, with probabilities P(*FEW*) and P(*MANY*).

soonish rather than at some point in the distant future when vastly many more people have been born. I emphasize "appear as if" because I think there is in fact no more than an appearance here; but more on this in a later section.

Sowers wants to avoid this implication by denying the second of the "facts" that were used in the deriving the above inequality, asserting instead that $Cr(e \mid MANY) = 1$.[3] If we do set the conditional credence of *e* given *MANY* equal to one, then the last inequality sign in the expression turns into an equality sign, and the ratio of the posterior probabilities of *MANY* and *FEW* becomes identical to the ratio of their priors. In other words, there is no probability shift in favor of *FEW*, no worrisome doomsday-like implication.

However, merely observing that the Doomsday argument can be blocked by setting $Cr(e \mid MANY) = 1$ is not enough to refute it. We must ask whether this move is justified. It seems clear that it is not.

To assert that $Cr(e \mid MANY) = 1$ is to assert that in the amnesia chamber, where *ex hypothesi* you are ignorant of your birth rank, you should nonetheless assign *all* your credence to *e*, the hypothesis that your birth rank is no greater than N_{FEW}. If you assign all your credence to that hypothesis, there is no credence left for the hypothesis that your birth rank is greater than N_{FEW}. You would therefore have to assign zero credence to that hypothesis. But that is unreasonable. In a situation where you have no direct information about your birth rank, where the next thing you learn might well be that *MANY* is true and that you are one of the people who have birth ranks greater than N_{FEW}, you should surely assign this contingency at least *some* credence. If you don't do that, you could never come to accept *e* through Bayesian conditionalization, no matter how much empirical evidence for *e* you may later accumulate.[4]

Against all this, what justification does Sowers offer for holding that $Cr(e \mid MANY) = 1$? He suggests that a

> way to see this point is to suppose that the evidence *e* consists of the statement 'my rank is *n*'. In the Doomsday situation this statement is equivalent to the statement 'there are at least *n* persons who will ever have existed'. (p. 43)

Since *MANY* implies that at least *n* persons have existed (assuming we interpret *e* as our actual evidence, so that $n=n_{birth}$), the conditional probability of *e* given *MANY* should be set equal to one.

This reasoning, however, rests on a faulty premiss. The claim that (in the doomsday situation or any situation) the statement "my rank is *n*" is equivalent to the statement "there are at least *n* persons who will ever have existed" is false. While "my rank is *n*" implies "at least *n* persons will ever have existed", the converse is not true. From at least *n* persons having existed it does not follow that my rank is *n*. (E.g. at least

---

[3] p. 43

[4] Such super-confidence in *e* could prove costly for those who are willing to demonstrate the sincerity of their stated beliefs by putting their money where their mouths are. A person who follows Sowers in setting $Cr(e \mid MANY) = 1$ and is willing to accept a bets on odds that according to her own professed view should bring her a positive expected payoff, not just in the amnesia chamber but also in various other situations of a similar kind that could be arranged, would open herself to systematic exploitation.

one person has existed, but I'm not the first person.) The statement that my rank is *n* is therefore *not equivalent to*, but rather *logically stronger than* the statement that at least *n* persons will ever have existed.

Ignoring the difference between these two statements, and focusing on the weaker of them, is disastrous when assessing the soundness of the Doomsday argument. For all evidence must be taken into account, if it has a bearing on the issue, and it is precisely the information about one's own birth rank that is meant to do much of the work in the Doomsday argument. The essence of the Doomsday argument is that we have useful indexical information in our knowledge of our own birth ranks, information that is alleged to have an unexpected bearing on what we should believe about the life-expectancy of our species.

Since the statements "my rank is *n*" and "at least *n* people will ever have existed" are not equivalent, the argument for setting $Cr(e \mid MANY) = 1$ in the amnesia chamber is unsound.

4. On the idea that only one sample-value is possible at any given time

It seems that the basic intuition driving Sowers' critique is that the Doomsday argument goes awry because the "sampling" that it invokes is of a cross-temporal and, allegedly, an illicit nature:

> What is crucial is that a correlation has been enforced … which renders the sampling process patently non-random … it is only possible to sample a seven after it has become impossible to sample a four and before it becomes possible to sample a ten. The strictures of random sampling, on the other hand, require that all results are equally probable at any time a sample is taken. (p. 40)

As I explained in the introductory section, the Doomsday argument does not insinuate that some objective sampling actually takes place; it says merely that we should in some ways assign credence *as if* such sampling had occurred. The plausibility of these credence assignments is what is at stake. It may be useful to look at a couple of thought experiments that both in some ways mirror the doomsday situation but that differ from one another precisely in regard to whether a time-correlation – at which the finger of blame should be pointed according to Sowers – is present.

> *The Incubator (synchronic version)*
> A dungeon contains 100 cells, numbered on the outside consecutively from 1 to 100. A machine, the "incubator", flips a coin. If the coin falls tails, the incubator does the following: It creates one observer in each of the first 10 cells. For half a day, these observers don't know which cells they are in; but in the second half of the day, they are shown the numbers of their cells. At the end of the day they are all killed. If the coin falls heads, then a similar procedure is followed except with a hundred observers (one in each cell) instead of only ten. No other observers exist than those in the dungeon, and everybody knows all of the above.

In this gedanken, no time-correlation exists between the sample-values and the time of the quasi-sampling. All sample values are determined at the same time and all observers exist concurrently (rather than in sequence as in the doomsday situation and the Amnesia chamber). Presumably, therefore, Sowers would have no objection to the credence assignment specified by applying the Self-Sampling Assumption. On this assignment, the people in the dungeon should have the following credences in the morning:

$$Cr(Heads) = Cr(Tails) = \tfrac{1}{2}$$
$$Cr(\text{"}I'm\ in\ one\ of\ cells\ \#1\text{–}\#10\text{"} \mid Tails) = 1$$
$$Cr(\text{"}I'm\ in\ one\ of\ cells\ \#1\text{–}\#10\text{"} \mid Heads) = \tfrac{1}{10}$$

From which is follows by Bayes' theorem that

$$Cr(Tails \mid \text{"}I'm\ in\ one\ of\ cells\ \#1\text{–}\#10\text{"}) = \tfrac{10}{11}$$
$$Cr(Heads \mid \text{"}I'm\ in\ one\ of\ cells\ \#1\text{–}\#10\text{"}) = \tfrac{1}{11}$$

We can contrast this with the following gedanken in which a time-correlation is enforced:

> *The Incubator (diachronic version)*
> As before, there is a dungeon with 100 consecutively numbered cells. The incubator flips a coin and does the following it falls tails: It first creates one observer in cell #1. For half a day, this observer doesn't know he is in #1, then he is shown the number of his cell, and after knowing this for the second half of the day, he is killed. On the second day, the incubator creates a new observer in cell #2. This observer is also ignorant about which cell she is in for half a day, then knows that she is in cell #2 for half a day, and is then killed. The process continues to cell #10, at which point the experiment ends. If the coin falls heads, a similar procedure is followed but for one hundred days. No other observers exist than those in the dungeon, and everybody knows all of the above.

This diachronic version of Incubator exhibits the same time-dependence as the doomsday situation and the Amnesia chamber. It is possible to find oneself in cell #7 only after it has become impossible to find oneself in cell #4 and before it becomes possible to find oneself in cell #10. Presumably, then, Sowers is committed to the holding that the Self-Sampling Assumption is inapplicable to this case and that credence should be assigned in some different way. In view of the rather close parallel to the Amnesia chamber, he might e.g. hold that a person who is alive in one of the first ten days should assign set the credence $Cr(\text{"}I'm\ in\ one\ of\ cells\ \#1\text{–}\#10\text{"} \mid Heads) = 1$, even before noon.

Yet treating the synchronic and the diachronic versions of Incubator differently seems to me undesirable, for in regard to certain central features they are very similar. In particular, although there is a time-correlation in the diachronic version between what day it is and which cell one is in, this correlation is unusable because both correlates are unknown. If one knew what day it was, one could infer which cell one was in, and vice

versa; but since one doesn't know either, one's knowledge that there is a correlation is to absolutely no avail. It doesn't help that one is able to determine which day it is as by saying that "the day is today", any more than one is helped in the synchronic version by being able to identify one's own cell as "this one". For what is needed is information about whether the current day is the first day, or the second day, etc., or alternatively, that one's cell is #1, or #2, and so on; and such information is not forthcoming until in the afternoon, in both versions of Incubator.

The idea that the "strictures of random sampling" require that all samples be equally probable at any time a sample is taken is a red herring. *There is no random sampling* in the situations that we have been discussing, or if there is, it is incidental. (We didn't specify whether the incubator distributed the people amongst the cells randomly or in some predetermined fashion. We didn't specify this because it didn't matter to the rational credence assignments for the people in the cells.) What there is, rather, is potential uncertainty about one's position in the world. There is also a suggestion that information about one's position may be epistemically linked to non-indexical hypotheses, and there is a proposal for how to assign credence under such conditions. This proposal invokes randomness as a heuristic but it does not claim that there is a corresponding set of physical chances or objective randomness. Therefore we can set aside scruples about physical randomness requiring availability of all alternatives at the time of sampling.

If a line is to be drawn between commonly accepted inference patterns and the reasoning embodied in the Doomsday argument, I must be drawn elsewhere than where Sowers wants it. In the following section, I shall try to indicate where I think it should be drawn.

5. Remarks towards a solution

This final section, in which I will try to say something about my own position, will by necessity be sketchy. This is because, in my view, a correct analysis of the Doomsday argument requires as a framework a theory of observation selection effects, and there is not space to expound such a theory here.

We can approach the problem by asking several questions regarding the Incubator thought experiment. Since to my mind, there is no significant difference between its synchronic and diachronic variants, the answers I shall suggest apply equally to both.

Question A. You find yourself in one of the cells one morning. What credence should you assign to Heads?

I say one half.

Others have argued that the answer should be greater than one half, invoking the Self-Indication Assumption, which states that the fact that you came into existence gives you some reason to favor hypotheses according to which many observers came into existence. Proponents of the Self-Indication Assumption would say that your credence in Heads should be $^{10}/_{11}$. I have argued against the Self-Indication Assumption elsewhere.[5]

[5] (Bostrom 2002)

Question B. You find yourself in one of the cells one morning. What conditional probability should you assign to being in a particular cell given Heads, or Tails?

I apply the Self-Sampling Assumption and get (where *i* is an integer between 1 and 100) the following recommendation for what people should believe in the morning:

$$Cr(I'm\ in\ cell\ \#i \mid Tails) = \begin{cases} 1/10 & for\ i \le 10 \\ 0 & for\ i > 10 \end{cases}$$
$$Cr(I'm\ in\ cell\ \#i \mid Heads) = 1/100\ .$$

The reasoning is simple. You know that if the coin fell tails then there are ten people, one in each of the first ten cells, any one of whom, for all you know, might be you. By the symmetry of the situation, you assign an equal credence to you being any of these ten people. The possibility of heads is handled in a parallel way.

The people in the various cells will presumably have somewhat different states of mind and experiences – one has an itch on her nose, another perhaps an ache in his knee. This would enable them to identify themselves not just demonstratively ("I'm *this* observer") but also by some definite description ("I'm the lady with the itchy nose"). However, they are unable to link such self-identifying descriptions to statements about what the number is of the cell they are in, so they cannot use this information about their idiosyncratic sensations to narrow down the space of possible locations at which they might be residing. The situation therefore appears to be analogous to one of a more familiar kind. Suppose that *x*% of the population has a certain genetic sequence *S* within the part of their DNA commonly designated as "junk DNA". Suppose, further, that there are no observable manifestations of *S* (short of what would turn up in a gene assay) and that there are no known correlations between having *S* and any observable characteristic. Then, quite clearly, unless you have had your DNA sequenced, it is rational to assign a credence of *x*% to the hypothesis that you have *S*. And this is so quite irrespective of the fact that the people who have *S* have qualitatively different minds and experiences from the people who don't have *S*. (They are different simply because all humans have different experiences from one another, not because of any known link between *S* and what kind of experiences one has.)

Question C. Later the same day you have discovered which cell you are in, let's say it is #7. What credence should you now assign to Heads?

My answer is that different credence assignments are possible, each of which seems rationally acceptable. Opinions of reasonable persons could differ on this point, reflecting difference in their prior credence functions. This point requires elaboration.

Based on my answer to question B, it could seem as if I were committed to answering question C by saying that Heads should be assigned a credence of $1/11$ and Tails a credence of $10/11$. For the above conditional probabilities,

$$Cr(I'm\ in\ cell\ \#7 \mid Tails) = 1/10$$
$$Cr(I'm\ in\ cell\ \#7 \mid Heads) = 1/100$$

together with $Cr(Heads) = Cr(Tails) = ½$ yield via Bayes' theorem that

$$Cr(Tails \mid I'm\ in\ cell\ \#7) = {}^{10}\!/_{11}$$
$$Cr(Heads \mid I'm\ in\ cell\ \#7) = {}^{1}\!/_{11}$$

Thus, what one gets if one conditionalizes the credences one has in the morning on "I'm in cell #7), is a posterior credence $^{1}\!/_{11}$ of Heads. How is an answer to C diverging from this result possible? Doesn't giving a different answer entail a violation of Bayesian conditionalization?

Let me try to briefly indicate, with a generous helping of hand waving, how one of the permissible answers to question C can be that Heads and Tails both be assigned a posterior credence of ½ and how this can be done without violating Bayesian conditionalization. Note, by the way, that assigning ½ credence to Heads and Tails in the Incubator thought experiment corresponds to *not* making any probability-shift in the doomsday situation, assuming the two cases are treated similarly.

If the sole difference in one's epistemic state between morning and afternoon were that at the latter time the state included the added piece of evidence that one is in cell #7, then Bayesian updating of one's morning-beliefs would preclude assigning Tails and Heads equal credence in the afternoon. However, there is another difference between the two epistemic states. In the morning, you know that "I am currently ignorant about which cell I am in", but of course you don't know that in the afternoon, since then you know that you are in cell #7. So in addition to gaining a piece of information (that you are in cell #7), you also *lose* a piece of information (that you are currently ignorant about which cell you are in).

Normally, this kind of loss of indexical information has no effect on one's posterior probabilities, being irrelevant to the hypotheses under consideration. In cases like Incubator, however, there is nothing to bar one from regarding the lost information as relevant. This is because the crucial questions in Incubator involve precisely the sort of indexical information that is being lost and gained.

To see this, we must first strengthen the Self-Sampling Assumption so that we get a principle that takes account of more indexical information: not only information about which observer one is, but also information about which temporal segment one (currently) is of that observer. To this end, we adduce the *Strong Self-Sampling Assumption*, which says that each observer-moment (i.e. each time-segment of an observer) should reason as if it were a random sample from all observer-moments in its reference class. (A defense of this principle and a detailed explanation of how it works are beyond the scope of this paper, but have been developed elsewhere.[6])

Now, observer-moments existing in the morning are in an important respect in an epistemically different situation from those existing in the afternoon. The former have no knowledge about which cell they are in and they might be actively pondering that very question. The afternoon observer-moments, by contrast, are under no uncertainty about which cells they inhabit. Moreover, afternoon observer-moments might also be in importantly different epistemic situations from one another: some of them might have found that they are in one of cells ##11-100; such observer-moments are no longer in any

[6] (Bostrom 2002)

doubt as to how the coin landed (they know it fell heads); while others, who have discovered that they are in cells ##1-10, will still be in suspense about the outcome of the coin toss.

Because of these differences, it seems one may elect – without undue arbitrariness – to place the morning observer-moments in a different reference class from the afternoon observer-moments, and to place those afternoon observer-moments who are still in doubt about the outcome of the coin toss in a different reference class from those (if there are any) who are now in a position to infer that the coin fell heads. If one does this, then the Strong Self-Sampling Assumption will generate different conditional credences for different classes of observer-moments.

The observer-moments existing in the morning will still assign the same conditional credences that we gave above. In particular, they will assign a ten times greater credence to being in cell #7 given Tails than given Heads:

$$Cr_{morning}(I'm\ in\ cell\ \#7 \mid Tails) = 10 \times Cr_{morning}(I'm\ in\ cell\ \#7 \mid Heads)$$

This is because, for a morning observer-moment, if the coin fell tails then the fraction of all observer-moments in its reference class who are in cell #7 is ten times greater than if the coin fell heads.

The observer-moments existing in the afternoon and who have found that they are in one of cells ##1-10, by contrast, will assign the same conditional credence to being in cell #7 given Tails as given Heads:

$$Cr_{afternoon,\ \#\#1\text{–}10}(I'm\ in\ cell\ \#7 \mid Tails) = Cr_{afternoon,\ \#\#1\text{–}10}(I'm\ in\ cell\ \#7 \mid Heads)$$

This follows directly from the fact that whether the coin fell heads or tails, the same fraction (namely, one tenth) of observer-moments in their reference class are in cell #7. We may note in passing that this would also hold if we had instead elected to place each afternoon observer-moment in a separate reference class of its own, perhaps on grounds that they each have different information about which cells they are in. With such a reference class definition, the fraction of all observer-moments who are in the same reference class as you (you being observer-moment who knows it is in cell #7) who are in cell #7 is one, independently of how the coin fell.

If there are any afternoon observer-moments who have found themselves in cells ##11-100, they would assign zero conditional credence to being in cell #7 whether given Heads or Tails, since the fraction of all observer-moments in *their* reference class who are in cell #7 is zero in either case. However, we need not concern ourselves with these observer-moments' beliefs for present purposes.

Our interest is focused rather on what an afternoon observer-moment who knows it is in cell #7 should believe. Given the reference class definition suggested above, such an observer-moment should assign a credence of ½ to Tails and an equal credence to Heads. For the prior credences of Heads and Tails were identical, and the observer-moment's conditional credences of being in cell #7 given Heads or given Tails are identical also. It then follows by Bayes' theorem that the posterior credences of Heads and Tails are the same.

Thus we see how we can get this alternative answer to question C, ½ rather than $\frac{1}{11}$, in two steps. First, we replace the Self-Sampling Assumption with a stronger principle and selecting an appropriate reference class definition. The Strong Self-Sampling Assumption takes more indexical information into account (not just information about which observer one is but also about which temporal part one of that observer is the current one) than the Self-Sampling Assumption, which it therefore trumps in cases of disagreement, since perfect rationality requires that all relevant information be taken into account. Second, we pick a suitable reference class definition, such as the one given above, which seems non-arbitrary and generally defensible. This gives the posterior credence of ½. (Bayesian conditionalization is not violated because the conditional credences we specified for the morning observer-moments were implicitly conditioned on the indexical information "I'm currently a morning observer-moment", which information is obviously not retained by the afternoon observer-moments.)

If we had instead picked a reference class definition that placed all observer-moments in Incubator in the same reference class, we would have recovered the original result, i.e. a posterior credence of Heads equal to $\frac{1}{11}$. Which of these answers we get depends on how we define the reference class. My suspicion is that commonly accepted constraints of rationality do not suffice to single out one reference class definition as the uniquely correct one. Instead, there may be room for reasonable people to adopt different reference class definitions. Since the definition of the reference class determines a way of taking indexical information into account, an element of subjectivity in the definition of the reference class corresponds to an element of subjectivity in the component of our prior credence functions that is concerned with indexical beliefs. That there should be such an element of subjectivity should not, I think, surprise us. Most acknowledge that there is some subjectivity in the choice of the non-indexical component in one's prior credence function. Once we consider it, it should be easy to accept that the indexical part may be similarly subjective.

To say that there is an element of subjectivity is not, of course, to say that *anything* goes, only that more than one thing goes. The desideratum of avoiding arbitrariness and the requirement that one's prior credence assignment be such as to enable reasonable scientific and common sense inductive practices impose considerable constraints on what can count as an acceptable reference class definition. These constraints become apparent when one extends consideration beyond one particular thought experiment, such as Incubator, to the full range of cases about which one has convictions.[7]

The approach that yielded the answer ½ to question C should transfer well to the doomsday situation, since the difference between the epistemic situations of the various observer-moments that were placed in different reference classes in this treatment of Incubator seem to be matched by at least as great differences in the epistemic situations between humans who live now and those who may live in the distant future if *FEW* is false. Answering ½ to question C corresponds to assigning a posterior credence to *FEW,* after taking one's birth rank into account, that is identical to the prior credence of *FEW*. (Naturally, this credence need not be one half but will instead be determined by the range

[7] See (Bostrom 2002) for a discussion of several other cases in both science and philosophy and of the constraints they impose on acceptable reference class definitions.

of normal empirical considerations that shape our beliefs humanity's prospects, such as the risks of nuclear war, germ warfare, asteroid strikes, etc., yet the reasoning is exactly the same.) If this is right, then the Doomsday argument is inconclusive: it does not rationally compel the probability-shift that the doomsayer asserts. At most, the Doomsday argument is persuasive to that subset of thinkers whose prior credence functions happen to have a special shape, such as the one that corresponding to placing all observer-moments in the same reference class.

References

Bostrom, N. (1997). "Investigations into the Doomsday argument." *Preprint*. http://www.anthropic-principles.com/preprints/inv/investigations.html

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

Sowers Jr., G. F. (2002). "The Demise of the Doomsday Argument." *Mind* **111**(441): 37-46.