WHAT, PRECISELY, IS CARTER'S DOOMSDAY ARGUMENT?

ABSTRACT. Paying strict attention to Brandon Carter's several published renditions of anthropic reasoning, we present a "nutshell" version of the Doomsday argument that is truer to Carter's principles than the standard balls-and-urns or otherwise "naive Bayesian" versions that proliferate in the literature. At modest cost in terms of complication, the argument avoids commitment to many of the half-truths that have inspired so many to rise up against other toy versions, never adopting posterior outside of the convex hull of one's prior distribution over the "true chance" of Doom. The hyper-pessimistic position of the standard balls-and-urn presentation and the hyper-optimistic position of naive self-indicators are seen to arise from dubiously extreme prior distributions, leaving room for a more satisfying and plausible intermediate solution.

1. INTRODUCTION

Anthropic reasoning principles leading eventually to one version of the so-called *Doomsday Argument* (see also Gott 1993, Nielson 1989) arose in the seventies in two papers by theoretical physicists (Collins and Hawking 1973, Carter 1974). Brandon Carter in particular is often credited as the most important early proponent of this sort of reasoning in general and the Doomsday argument in particular.

Leslie (1989, 1992, 1996), "working only from rumours about how Carter was running it", proposed a balls-and-urn version of the Doomsday argument. Bostrom (1999, 2001) presents similar "nutshell" cases. The following is representative:

Doomsday: Assume two equally likely scenarios: humanity will suffer extinction sooner (*Quick Doom*), in which case there will be a total of 200 billion humans, or humanity will suffer extinction later (*Later Doom*), in which case there will be a total of 200 trillion humans. You learn first that you are a member of this indeterminately sized population. At this point your credence in *Quick Doom* is $\frac{1}{2}$. Next, you learn that you are among the first 200 billion humans. That fact, conditional on *Quick Doom*, has probability one. Conditional on *Later Doom*, it has probability $\frac{1}{1000}$. Therefore, you ought to update your credence in *Quick Doom* to $\frac{1000}{1001}$ by, e.g., Bayes' Theorem:

$$P(Q|E) = \frac{P(Q)P(E|Q)}{P(Q)P(E|Q) + P(L)P(E|L)} = \frac{\frac{1}{2} \cdot 1}{\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{1000}} = \frac{1000}{1001}.$$

(Here Q = *Quick Doom*, L = *Later Doom* and E = *Early Birth Rank*.)

Proceeding from the assumption that this is his argument, it would be very easy to get the impression (as many have) that Carter is badly mistaken. One might reach this conclusion, for example, by considering an augmentation in which for each "civilization" in an infinite cosmos, a fair coin is tossed once the population reaches 200 billion. If *heads*, the civilization is destroyed. If *tails*, the civilization is destroyed when the population reaches 200 trillion. The mistake in **Doomsday**, one then reasons, is that at the point where you have learned you are human, but have yet to learn your birth rank, your credence in *Later Doom* should be exactly one thousand times greater than your credence in *Quick Doom*, for the reason that (in any large enough region) one expects that one thousand times fewer of your epistemic counterparts' civilizations suffer early extinction than do not.[1]

But **Doomsday** is an oversimplification–an extreme case of what Carter (1983) actually describes.[2] I don't claim that Leslie and Bostrom are confused on this point; each makes numerous qualifications and offers scenarios in which the Bayesian calculation of **Doomsday** is taken not to apply. Neither, however, offers an argument that is both general enough to avoid the need for qualification and formal enough to satisfy skeptics who insist that the favoring of *Quick Doom* be rendered in *some* sort of "rigorous calculation".[3] Our purpose here is to show what such a calculation looks like. This will expose, in particular, that Carter wasn't wrong at all–not, at least, for any of the reasons these skeptics have offered.

2. Carter's anthropic principle

Carter (1983) contains a concise formulation of the anthropic principle:

> "In a typical application of the anthropic (self-selection) principle, one is engaged in a scientific discrimination process of the usual kind in which one wishes to compare the plausibility of a set of alternative hypotheses, $H(T_i)$, say, to the effect that respectively one or other of a corresponding set of theories $T_1, T_2, \ldots$ is valid for some particular application in the light of some observational or experimental evidence E, say. Such a situation can be analysed in a traditional Bayesian framework by attributing *a priori* and *a posteriori* plausibility values (i.e. formal probability measures), denoted by p_E and p_S, say, to each hypothesis respectively before and after the evidence E is taken into account, so that for any particular result X one has
>
> $$p_E(X) = p_S(X|E),$$
>
> the standard symbol $|$ indicating conditionality. According to the usual Bayesian formula, the relative plausibility of two theories A and B, say, is modified by a factor equal to the ratio of the corresponding conditional *a priori* probabilities $p_S(E|A)$ and $p_S(E|B)$ for the occurrence of the result E in the theories, i.e.
>
> $$\frac{p_E(A)}{p_E(B)} = \frac{p_S(E|A)}{p_S(E|B)} \frac{p_S(A)}{p_S(B)}. \quad (1)$$"

[1] Such "self indication" or "double shift" reasoning is employed by Dieks (1992), Bartha and Hitchcock (1999) and others in direct response to **Doomsday**, and appears to be closely aligned with the majority thirder response to the so-called "Sleeping Beauty" problem. Bostrom's (1999) thought-experiment (the *Presumptuous Philosopher*) cautions against its too-liberal application, and is the most effective apologetic for Carter's methods I know of.

[2] By analogy with other cases of anthropic reasoning; he didn't publish on Doomsday *per se*.

[3] Leslie (1996) has on this point suggested that there isn't *need* of a formal mathematical presentation, but decades of entrenchment and thousands of pages of spilt ink suggest otherwise.

Carter stresses that the "Selected" or "Subjective" probability function p_S in (1) is related to an "Original" or "Objective" probability function p_O by $p_S(\cdot) = p_O(\cdot|S)$, "where S denotes the totality of all the selection conditions that are implied by the hypothesis of application of the theory to a concrete experimental or observational situation, but which are not necessarily included in the abstract theory" on which p_O is based. In all three of the examples discussed in Carter (1983), one has

$$\frac{p_S(E|A)}{p_S(E|B)} \neq \frac{p_O(E|A)}{p_O(E|B)};$$

indeed, this is the hallmark of anthropic reasoning as Carter understands it.

It's worth examining Carter's explanation for this in the first of these examples, which he takes to be "the classic example of an argument based on the anthropic principle". Here A is the hypothesis to the effect that the development of life is of common occurrence on 'habitable' planets, B is the hypothesis that life is very rare, even in favorable conditions, and E is the evidence consisting of the fact that on the only obviously 'habitable' planet we have yet been able to observe, namely our own, life does indeed exist.

> "If future astronomical progress should one day enable us to observe a second example of occurrence of life on a randomly chosen 'habitable planet' belonging to a not too distant star in our Galaxy, the corresponding *ab initio* probability ratio, $\frac{p_O(E|A)}{p_O(E|B)} >> 1$, would justify the induction that hypothesis A (that life is common) was the most likely. However, so long as the only example at our disposal is our own, no such inference is permissible, since the anthropic selection principle ensures, as a virtual tautology,. that one of the *a priori* conditions, S, that must be satisfied by the first planet available for investigation by us must be the prior occurrence of life, namely our own. Thus as in the previous example we obtain not only $p_S(E|A) = 1$ but also $p_S(E|B) = 1$, so that our observation has no discriminating power at all, and both...A and B remain equally viable."

In a second example B is the hypothesis to the effect that gravitational coupling strength is fixed across time, while the evidence E is that of a seemingly fortuitous mathematical relationship between the Hubble time and the gravitational coupling constant. (Hypothesis A is that coupling strength increases across time to preserve this relationship.) According to Carter, in this case $p_O(E|B) << 1$ since the relationship can hold only in one particular epoch under the hypothesis that the coupling strength doesn't vary, but $p_S(E|B) \approx p_S(E|A) = 1$ since "biological systems based on the same principles as our own" won't exist in times where the relationship doesn't hold. That is, the anthropic selection principle ensures that the seemingly fortuitous relationship between Hubble time and the gravitational coupling constant must be observed. As in the first example, then, hypothesis A is not confirmed by the observation.

In the third and final example, A is the hypothesis to the effect that the expected average time $\bar{t}$ intrisically most likely for the evolution of a system of observers intelligent enough to comprise a scientific civilization such as our own is geometrically small relative to the main sequence lifetime τ of a typical star, during which the energy output can maintain favorable conditions for life; hypothesis B is that $\bar{t}$ is geometrically large relative to τ. Now E is the evidence that the time $t_e \approx 4$ billion years necessary for the evolution of intelligent life on Earth is on the same order of magnitude (i.e. geometrically comparable) as the estimated main sequence lifetime $\tau_0 \approx 10$ billion years of the Sun. In this case, Carter would have us accept that $p_0(E|A)$ and $p_0(E|B)$ are both very small (and plausibly near each other). On the other hand $p_S(E|B) \approx 1$ (so that in particular B is confirmed at the expense of A by the observations), as Carter explains in this passage:

> "...the observation that t_e is comparable with the upper limit τ_0 is just what would be expected if we adopt the alternative hypothesis that the intrinsically expected time $\bar{t}$ is much longer than τ_0: in this case self-selection ensures that ours must be one of the exceptional cases in which evolution has proceeded much faster than usual; (...) there is no particular reason why we should belong to the even more exceptional cases in which evolution proceeds even more rapidly although, with the assumption that the Universe is infinite, such cases must of course exist."

Note two features common to these examples.

First, in all of these cases $\frac{p_S(E|A)}{p_S(E|B)} < \frac{p_O(E|A)}{p_O(E|B)}$; because observations must occur from the first person perspectives of life forms in some respects similar to us, they are systematically predisposed to favor, conditional on B, positions, situation or scenarios consistent with E–despite the fact that positions, situations or scenarios consistent with E might be comparatively rare conditional on B. (For A as well, perhaps, but relatively more dramatically for B.) For this reason one should not, despite this comparative rarity, discredit B on the basis of an observation of E. After all, that the first observation made would be consistent with E is exactly what one should expect conditional on B, regardless of what vast volumes of spacetime one might have to scour in order to locate such an observation.

Second, the ratio $\frac{p_S(A)}{p_S(B)}$ is assumed to be equal to $\frac{p_O(A)}{p_O(B)}$. That is to say, the concrete "selection conditions" aren't taken to favor the hypothesis A over the hypothesis B on the basis that "more" (in the sense of density if not actual numbers, in the apparently default case that the Universe is assumed infinite conditional on either hypothesis) observers are predicted conditional on A than on B. Here is where it is essential that A and B are "theories", i.e. families of probability laws on the set of complete Universe trajectories. Note that if, to the contrary, A and B were chance events or ineliminably indexical assertions (*heads* and *tails* or *Quick Doom* and *Later Doom*, for example), this assumption would run counter to frequentist views of credence.[4] As it stands, the assumption is consistent with the majority

[4]Compare Sleeping Beauty or the objections to **Doomsday** in Section 1 related to sequences of civilization in an infinite cosmos. Note however that Bostrom (1999) and possibly Leslie believe that whether such frequentist reasoning is appropriate in **Doomsday** is sensitive to the *actual* presence of said "outsiders", i.e. other civilizations. Since Carter seems unconcerned that the infinitude of the Universe might affect his analyses, I have assumed that he would regard this issue as a red herring, and be sympathetic to the "theory" distinction alone. Compare also Antony Lewis's *SIA-C* (Lewis 2001), which applies "within different probabilistic outcomes of a correct theory", does not favor "wrong theories with a larger number of observers" (i.e. isn't

response (no skewing of prior credences in favor of theories associated with greater expected population) to Bostrom's "Presumptuous Philosopher" experiment.[5]

3. An adequate formalization of Carter's Doomsday argument

Before proceeding to Carter's version of the Doomsday argument, I shall give a brief example showing how (1) is to be used. Consider a coin whose behavior is known to be correctly described by one of two theories. The first theory is "the probability that this coin lands *heads* on any particular toss is $\frac{1}{2}$, independently of how it lands on other tosses" (call this theory $T_{1/2}$). The second is "the probability that this coin lands *heads* on any particular toss is $\frac{1}{3}$, independently of how it lands on other tosses" (call this theory $T_{1/3}$). Say that, initially, we are indifferent as to which theory is true. That is, $p_S(T_{1/2}) = \frac{1}{2} = p_S(T_{1/3})$.

Suppose next that we observe E, i.e. that the coin presently lands *heads* when tossed. By (1), we have:

$$\frac{p_E(T_{1/2})}{p_E(T_{1/3})} = \frac{p_S(E|T_{1/2})}{p_S(E|T_{1/3})}\frac{p_S(T_{1/2})}{p_S(T_{1/3})} = \frac{(1/2)}{(1/3)}\frac{(1/2)}{(1/2)} = \frac{3}{2}.$$

Then our posterior credences in the two theories under consideration are $p_E(T_{1/2}) = \frac{3}{5}$ and $p_E(T_{1/3}) = \frac{2}{5}$, respectively.

Note that, by virtue of affecting one's credences in rival theories, the present observation of a chance event (such as *heads*) may affect one's credence in another chance event. To return to our example (where E is the event that a present toss lands *heads*), let F be the event that a subsequent toss will land *heads*. Then

$$p_S(F) = p_S(T_{1/2})p_S(F|T_{1/2}) + p_S(T_{1/3})p_S(F|T_{1/3}) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{3} = \frac{5}{12},$$

whereas

$$p_E(F) = p_E(T_{1/2})p_E(F|T_{1/2}) + p_E(T_{1/3})p_E(F|T_{1/3}) = \frac{3}{5}\cdot\frac{1}{2} + \frac{2}{5}\cdot\frac{1}{3} = \frac{13}{30}.$$

We are now ready to proceed. In the following, I shall restrict application of (1) to the only sort of hypotheses implicitly sanctioned by Carter's own practice: hypotheses to the effect that one or another competing *theory* is true. In order to avoid straying too far away from the format of **Doomsday** (which might complicate comparison), I shall carry over the assumption that for human-like species there are only two equally (from an "outside" perspective) likely possibilities: *Quick Doom* (population 200 billion) and *Later Doom* (population 200 trillion).

"Presumptuous") and agrees with "the frequentist probability" if you "make many universes"– e.g. countably iterate a finite Universe under a single theory.

[5]I would urge more entrenched proponents of rival anthropic practices to grant some leeway here...perhaps view the whole paper as more descriptive than normative. (Though I do happen to believe that Carter is right, there's a reason I didn't entitle the paper "Why Carter's Doomsday Argument is Right".) I'm mainly trying to make it clear, by looking at a succession of examples, that Carter only employs (1) in cases where A and B are competing *theories* (as I've defined this notion); in particular, in cases in which one has no actual or nomologically possible counterparts for whom A and B have different truth values than they actually have. I don't claim to be making any advance on the "Presumptuous Philosopher" intuitions that this practice is "correct".

The theories grounding the competing hypotheses for consideration in (1) that I shall consider are given by $\{T_x : x \in [0,1]\}$, where T_x is the assertion that the expected (and almost sure, should the Universe be infinite) density[6] of *Quick Doom* in human-like species, *conditional on the actual complete theory*, is x. Of course for a unique $x_a \in [0,1]$, T_{x_a} is the "correct" theory. Since the agent judges *Quick Doom* and *Later Doom* to be equally likely for any given human-like species, the expectation of x_a (from an "outside" perspective) is $\frac{1}{2}$. On the other hand one should allow for the agent's distribution for x_a to be potentially continuous and quite diffuse, for one cannot necessarily anticipate the extent to which the agent will be able to discredit, from first principles, potential (regions of) values for x_a. Denote the probability density function for this distribution by $g : [0,1] \to [0,\infty)$.

Now according to (1), upon observation of E = *my birth rank is at most 200 billion*, one ought to multiply the density function g by a factor proportional[7] to

$$p_S(E|T_x) = \frac{200,000,000,000x + 200,000,000,000(1-x)}{200,000,000,000x + 200,000,000,000,000(1-x)} = \frac{1}{1000 - 999x}.$$

That is to say, the agent's posterior density function for x_a will be

$$h(x) = \frac{kg(x)}{1000 - 999x}, \text{ where } k = \Big(\int_0^1 \frac{g(x)}{1000-999x}\, dx\Big)^{-1}.$$

Posterior credence in *Quick Doom* is now just posterior expectation of x_a, i.e.

$$p_E(D_1) = \int_0^1 xh(x)\, dx.$$

Examples:[8]

1. Let $g(x) = 6x(1-x)$. Then $h(x) \approx \frac{6x(1-x)}{.0029674(1000-999x)}$ and

$$p_E(D_1) = \int_0^1 xh(x)\, dx \approx \int_0^1 \frac{6x^2(1-x)}{.0029674(1000-999x)}\, dx \approx .663683.$$

2. Let $g(x) = 1$, i.e. the uniform prior. Then $h(x) = \frac{999}{\ln(1000)(1000-999x)}$ and

$$p_E(D_1) = \int_0^1 xh(x)\, dx = \int_0^1 \frac{999x}{\ln(1000)(1000-999x)}\, dx \approx .856236.$$

[6]For those who might be squeamish about densities, substitute "ideal subjective probability of" (still conditional on the actual complete theory).

[7]A referee: "Perhaps you're just using Bayes Theorem here, not (1)." It's (1). Note in particular that $\frac{h(x)}{h(y)} = \lim_{\epsilon\to 0} \frac{p_E(T_{[x,x+\epsilon)})}{p_E(T_{[y,y+\epsilon)})} = \lim_{\epsilon\to 0} \frac{p_S(E|T_{[x,x+\epsilon)})}{p_S(E|T_{[y,y+\epsilon)})} \frac{p_S(T_{[x,x+\epsilon)})}{p_S(T_{[y,y+\epsilon)})} = \frac{g(x)}{g(y)} \frac{p_S(E|T_x)}{p_S(E|T_y)}$ a.e.

[8]There is no implicit claim that the below distributions have direct relevance to our problem. They are simply common distributions (1 and 4 are Dirichlet distributions, 2 and 3 are uniform distributions of two different sorts) exhibiting a variety of concentration patterns. We're calling attention to the dependence of *Quick Doom*'s posterior credence on the pattern of concentration.

3. Let $g(x) = k(\ln(1-x) - \ln x)$ for $0 < x \leq \frac{1}{2}$, $g(x) = k(\ln x - \ln(1-x))$ for $\frac{1}{2} < x < 1$, where $k \approx 1.386284$. Then $h(x) \approx \frac{g(x)}{.0184138(1000-999x)}$, and

$$p_E(D_1) = \int_0^1 xh(x)\,dx \approx \int_0^1 \frac{xg(x)}{.0184138(1000-999x)}\,dx \approx .946642.$$

4. Let $g(x) = \frac{1}{\pi\sqrt{x(1-x)}}$. Then $h(x) = \frac{10\sqrt{10}}{\pi\sqrt{x(1-x)}(1000-999x)}$ and

$$p_E(D_1) = \int_0^1 xh(x)\,dx = \int_0^1 \frac{10\sqrt{10}x}{\pi\sqrt{x(1-x)}(1000-999x)}\,dx \approx .969347.$$

Though all non-singular distributions favor *Quick Doom* to some degree, this argument will never violate our intuitions by adopting a posterior credence outside the convex hull of our prior support for the true probability of *Quick Doom*.[9] As to which distribution is apt, there is anecdotal evidence that Carter (1983) would opt for one concentrated toward the extremes. To wit:

> "...the very complicated mechanisms governing the evolution of living systems cannot yet be analysed, still less predicted, in other than very vague qualitative terms. We certainly do not know enough to predict from first principles whether the expected average time $\bar{t}$ which would be intrinsically most likely for the evolution of a system of 'intelligent observers', in the form of a scientific civilization such as our own, should take much less or much more time than is allowed by the external restraints that limit the duration of favourable conditions. In such a state of ignorance, both of these two alternative possibilities should therefore be retained for consideration as not implausible a priori. Only the intermediate borderline case, in which the intrinsically most likely evolution time came out to be of just the same order as the time allowed by external restraints, could be set aside in advance, as being much less plausible *a priori*..."

Reasoning similarly in the current case, one arrives at a posterior credence in *Later Doom* of perhaps just a few percent, as in the latter two examples above.

4. Is Carter's argument sound?

Assuming that what I presented in the previous section really is Carter's argument, the question arises as to its soundness. There are grounds for believing the argument to be sound. If one's anthropic reasoning is constrained by both frequentism (e.g., the thirder response to Sleeping Beauty) and Bostrom's intuitions about the Presumptuous Philosopher (forcing the "event/theory" distinction) one would find it difficult, in reasoning anthropically, to do other than respect Carter's practice. We, for example, came to this conclusion prior to having read Carter; this is just what appears to have been forced by the most plausible intuitions.

[9] As the expectation of ideal subjective probability, one's credences should never fall outside the support of ideal subjective probability's distribution; Carter's anthropic principle teaches that while selection effects may alter this distribution, they ought never to expand its support.

Those who have resisted frequentist intuitions (David Lewis and the so-called "double halfers") in anthropic reasoning have been derided mercilessly in the literature, and I shall do no piling on in these pages. What we'll consider instead are two attempts by frequentists (thirders and double shifters, roughly) to oppose the event/theory distinction. These arise in the context of Bostrom's Presumptuous Philosopher thought experiment. Recall that, in this experiment, we are asked to imagine that there are two (and only two) competing "theories of everything", namely T1, on which there will be an expected trillion trillion observers in the cosmos, and T2, on which there will be an expected trillion trillion trillion observers in the cosmos. It would be presumptuous (says Bostrom) to say that T2 is overwhelmingly likely to be the true theory solely on the basis of anthropic reasoning favoring theories that give rise to greater populations.[10]

Kenneth Olum, like others employing the double shift, is encouraged that "treating possible observers in the same way as those who actually exist" skirts **Doomsday**. And, he is willing to follow the presumptuous philosopher in thinking of observers under an incorrect theory as "possible".[11] So accepting Bostrom's priors in the thought experiment, Olum bites the bullet and concedes that posterior probability in T1 ought to be miniscule. However, Olum seeks to avoid any embarrassment from this practice by holding that Bostrom's priors are unrealistic:

> ...it is possible that one should think that a theory involving a very large universe is unlikely in proportion to the size of the universe it proposes. In this case, the presumptuous philosopher is wrong, because the tiny *a priori* probability for the theory with the larger universe cancels out its larger number of observers.

Note that Olum's position suffers if his epistemic distribution over the number of observers is infinite. For if we replace T2 by the union T3 of theories on which "the number of expected observers is *at least* a trillion trillion trillion" and one's prior distribution assigns T3 any positive probability and the expected number of observers is infinite, Olum's posterior assigns T3 probability 1.

The idea of a finite expectation distribution has some attractiveness, but ultimately I don't see how, conditional on the universe (i.e. "everything") being finite, one might reasonably assign even a countably additive distribution to the number of observers, much less one having finite expectation. I don't see how, in particular, that there could be an N such that one had 99.9999999999% confidence that there would be at most N observers, given that every N is miniscule relative to many presumably reasonable possibilities for the expected number of observers in a "vast" universe.[12] So I doubt Olum is on the right track here.

[10]Note however that Bostrom too fails to pick up on the event/theory distinction, writing (Bostrom 2002) "It is hard to see what the relevant difference is between this case and [fair coin-based balls-and-urns experiments]."

[11]Carter would likely disagee (for the relevant sense of "possible). A "theory of everything", on this view, is not a matter of contingency. If you think it is, this attitude says you are giving "everything" too narrow a scope...it should encompass what you are now thinking of as counterfactual universes, if the theory deems that these might have been actual.

[12]Anyone who believes that the universe might be finite of course faces other hard questions. How would it bootstrap itself into existence? And if it did so once, what would stop it from doing

Another way to deny the force of the Bostrom's thought experiment is to grant that the presumptuous philosopher is right to make the shift that he does in favor of greater populations, but then neglects a second shift...a "Doomsday" type shift...back in favor of the less populous scenario. Bostrom had explicitly denied this move in his (2002):

> So why cannot DA be applied in The Presumptuous Philosopher to cancel the SIA-induced probability shift in favour of T2? The answer is that in the absence of knowledge about our absolute birth ranks, DA works by giving us information about what fraction of all species are short-lasting. (That we should be at an "early" stage in our species is more likely, according to the DA-reasoning, if a large fraction of all observers find themselves at such an early stage i.e., if long-lasting species are rare.) This information about what fraction of all species are short-lasting (a larger fraction than we had thought) in turn tells us something about our own fate (that it is more likely that we are a short-lasting species). But it does not tell us anything about how many species, and thus about how many observers there are in total.

Bradley Monton (2003) took issue with the above passage from Bostrom, arguing that you can, after all, run the Doomsday argument without knowledge of birth rank. If so, it would allow double shifters to flout the event/theory distinction and apply self indication reasoning even with respect to theories (and so resist Carter's argument). Monton writes "What is needed is a property p such that you know you have p, and the total number of observers expected to have p would be the same regardless of (whether T1 or T2 is true)." There are such properties. You could, for example, use your life history to generate, in some canonical fashion, a random number x uniformly distributed on the set $\{1, 2, \ldots, 10^{12}\}$. You now let p be the property "Either I live in a T1 universe or I live in a T2 universe and my life history generates x." You know yourself to possess p, and the expected number of observers possessing p is a trillion trillion on either theory.

But this property clearly won't do the job, because you also know yourself to possess the stronger property "I live in a T1 or T2 universe, and my life history generates x." Monton requires that you should necessarily know that the strongest property p^* you know yourself to possess should be instantiated equally frequently (in expectation) on each theory; but this is (mathematically) impossible, because if we let $E_i(p^*)$ be the expected number of observers for whom p^* is their strongest known property conditional on T_i, then $\sum_{p^*} E_i(p^*) = 10^{12+12i}$. It follows that the expected number of persons having a strongest known property p^* for which $E_1(p^*) = E_2(p^*)$ is at most 10^{24}. So if T2 is the correct theory and you are an observer selected uniformly at random, the probability that your particular p^* satisfies $E_1(p^*) = E_2(p^*)$ is at most one in a trillion. So Monton's argument fails.

so again? Maybe it should come as a surprise that there's anything at all, but given that there obvious is something, the most natural thought is that there's just everything, infinitely often, and the correct theory of the whole mess is just that which ascribes to each finite constellation of "somethings" the almost sure asymptotic frequency of its instantiations.

5. Even **Doomsday** presentors think the true argument is safer

It's clear that others have unboxed Carter's argument in the manner of Section 3. Lewis (2001), for example, formulated rather precisely the model of anthropic reasoning on which that argument rests. What's perhaps more surprising is that there is textual evidence that both Leslie and Bostrom essentially (when push comes to shove) think of the Doomsday argument in the manner of Section 3 as well. One therefore wonders why balls-and-urns versions have survived at all.[13]

In this section we'll observe how Leslie and Bostrom, when pushing back against the double shift, often retreat to safer positions more in line with Carter's practices. The most prominent instance we've seen (the Presumptuous Philosopher, where Bostrom makes population size a matter of theory rather than chance).

Here's a more explicit case from (2002, p. 114), where Bostrom has just finished acknowledging that the Dieks double shift is sound in a case where we can be certain that there are equal, large numbers of large and small actual populations:

> It is worth emphasizing, however, that suspecting that there are extraterrestrial civilizations does not damage DA if we don't have any information about what fraction of these alien species are "small". What DA would do in this case (if the argument were sound in other respects) is give us reason to think that the fraction of small intelligent species is greater than was previously held on ordinary empirical grounds.

What Bostrom is talking about is confirmation, by low birth rank, of theories on which greater fractions of intelligent species are small, i.e. Carter's true argument. But since in realistic cases we obviously do have an absence of knowledge about our absolute birth ranks, it seems that this just *is* the Doomsday argument.

Next we see how Leslie (1992) responds to the double shift threat to **Doomsday**:

> (Dieks) believes that Carter's argument commits a dreadful error: it forgets that there would be *enormously many more humans making their observations* if the human race lasted for enormously long. (...) If his argument works then it delivers the curious result that we should be not in the slightest influenced by finding ourselves in the 1990s. For a human to find himself or herself there would be not particularly improbable on the hypothesis of Quick Doom; it could be very, very improbable on the hypothesis of Doom Deferred; and yet, says Dieks, this is absolutely nothing in favour of the one hypothesis rather than the other. But is this not too like arguing that finding oneself bitten by a dog gives one absolutely no evidence for the hypothesis that the dog habitually bites?

Recall our explanation (at the beginning of Section 3) of how seeing a coin land *heads* affects one's credences in various hypotheses (*theories*) as to the coin's tendency to so land. Analogously, seeing Leslie's dog bite affects one's credences in various hypotheses as to the dog's tendency to so act. It isn't chance events or

[13] As recently as March 2019 Bostrom presented the balls-and-urns version of Doomsday on a podcast (1:14 of https://samharris.org/podcasts/151-will-destroy-future/)–whilst stressing that he does not endorse the argument. The true argument can be endorsed without embarrassment.

any other ineliminably non-indexical assertions that receive direct confirmation (such events may receive indirect confirmation transmitted through the theories, as we described in the case of the coin) here but hypotheses ("theories") about the long-run frequency of biting incidents in that particular dog's encounters with humans. So this appeal to the dog's biting habits is a subtle retreat to Carter's original principles–a retreat that gets far less subtle in the very next paragraph.

> Consider L.S. Marochnik's theory that in our galaxy it is *only in a narrow band* that planets have a reasonable chance of forming, and observers of evolving. (The band is at a distance from the galactic centre at which a density wave moves at roughly the same speed as the stars.) Our actually finding ourselves inside this band should surely count as *some* evidence in favour of such a theory.... Yet Dieks's reasoning suggests that if observers were instead to be found throughout the galaxy, then the correspondingly richer opportunities of finding oneself in the galaxy *would exactly compensate for* the lesser chance of finding oneself in the band in question.

That the double shift doesn't apply here is straight up Carter. For obviously Marochnik's hypothesis isn't only about *our* galaxy. In fact the double-shift reasoning would be appropriate if we were split on whether Marochnik's hypothesis was true of our galaxy because we were convinced that half of all galaxies are "Marochnik galaxies" and the other half were teeming with Earth-like planets. (In such a case, we should think, prior to conducting the measurement showing our own position, that we were more likely to be in a galaxy teeming with Earth-like planets.) But Marochnik's hypothesis isn't of that nature. It's what we've been calling a *theory*; so if we're split on whether it holds true of our galaxy, that's because we're split on whether it's true of all galaxies, or false of all galaxies.

Some of the most interesting sections of Leslie (1996) are sections in which the double shift is essentially capitulated to. In the following passage Leslie pins the blame for such capitulation on "radical indeterminism":

> There is always a firm fact of the matter...whether or not we know it, of how many names there actually are in any given urn. But the world may be...*a radically indeterministic world* in which there isn't yet any relevantly similar fact of how long the human race is going to last. However the doomsday argument does need such a fact in order to run smoothly. (...) Still, it only reduces the power of the Doomsday argument instead of destroying it. (...) If our world were indeterministic, there could well be no usable fact of how long the human race will last. Urn analogies might therefore work only rather poorly. (...) Even so, the doomsday argument would retain considerable force...

Leslie is speaking, unmistakably if elliptically, about running Carter's true argument in a case where a balls-and-urns version fails to "run smoothly" owing to "radical indeterminism" in the process by which the urn is selected. What we would say here is that Leslie is on the right track, but conflates *deterministically true* with *necessarily true*. Determinism says that everything follows from the initial conditions...it doesn't say that initial conditions aren't contingent, and the effect Leslie is noting here is no different if all of the contingency is packed into the initial conditions. Carter's principles require competing theories that are necessarily true where they are true and necessarily false where they are false...it isn't sufficient that they be deterministically true where they are true and deterministically false where they are false. These hypotheses must be such that their truth values remain constant over counterfactual cases. This is true of deterministic propositions so long as one fixes initial conditions across counterfactual cases, but there aren't any grounds for fixing initial conditions across counterfactual cases. If the universe is deterministic and finite but its initial conditions are contingent (subject to probabilistic description, even), Carter's principles admit counterfactual agents inhabiting universes with different initial conditions (but no counterfactual agents inhabiting universes operating under different theories) "into the reference class". (And if the universe is infinite, one may dispense with the encumbrance of counterfactuals entirely, since the space of actual counterparts is sufficiently rich and has the right statistics almost surely.)

And one mustn't forget the "competing" aspect in "competing theories". In order for Carter's true argument to induce a shift toward more likely early doom, there has to be uncertainty as to the correct theory, and in particular as to the ideal epistemic chance of early doom. Models involving dice that are known to be fair won't exhibit this effect, so it's important to raise the possibility of loadedness in any Doomsday-supporting experiment with dice. (The double shift applies in models with fair dice.) Leslie (1996, p. 244) does so in the following passage:

> You find yourself in a room crammed with as many humans as will have been born in the real world before AD 2150. (...) Although impressive in size, the room is tiny by comparison with a second room, a room large enough to contain ninety thousand times more people. Its ever actually containing them was to depend on God's throw of two [radically indeterministic] dice.... Just if the dice fell appropriately [not double six], God would create all those hugely many other people too. (...) Suppose...further...that you've strong reasons to suspect the dice were loaded (which nicely reflects our frequent sense of insecurity when estimating probabilities) (...) It's in the small room that you actually find yourself, however. All things considered, therefore, it would seem that God's dice had probably landed [double six]"

Leslie stresses that the argument depends crucially on the loadedness of the dice:

> Once again you find yourself in the small room. (...) Let's specify, just as before that God's dice are radically indeterministic. (...) Let's say you know this, adding that you even know for sure that the dice are utterly fair, unloaded. (...) Can finding yourself in the small room give you any grounds for picturing the dice as landing double-six? The answer is that it cannot. The dice are fair.... The probability that they will land double-six is therefore exactly one in thirty-six. (...) Does this destroy the doomsday argument? Unfortunately it only weakens it. In the world as it actually is, we could have no assurance that the probability that the human race would survive for this or that number of years was precisely such and such.

So in this passage Leslie explicitly capitulates to the double shift when one knows the true chances, at least when one also has "radical indeterminism...but, again, I would say–as would Bostrom (see footnote 9 on p. 108 of Bostrom 2002)–that Leslie is mistaken to think that anything turns on this feature. He then falls back on Carter's true argument in a case where one *doesn't* know the true chances.

It seems clear, then, that the double shift is a formidable enough obstacle for **Doomsday** that its leading popularizers invariably retreat from it to the well trodden ground of Carter's principles. The best explanation for this is the obvious one. Namely, that while there are obvious and probably fatal problems with the balls-and-urns versions, Carter's true argument remains "alive and kicking".

References

Bartha, P. and C. Hitchcock (1999). No One Knows the Date or the Hour: An Unorthodox Application of Rev. Bayes's Theorem. *Philosophy of Science (Proceedings)* 66:S229-S353.

Bostrom, Nick. 1996. Investigations into the Doomsday Argument. Online. http://www.anthropic-principle.com/preprints/inv/investigations.html

Bostrom, Nick. 1999. The Doomsday Argument is Alive and Kicking. *Mind* 108:539-553.

Bostrom, Nick. 2001. The Doomsday Argument, Adam & Eve, UN^{++}, and Quantum Joe. *Synthese* 127(3):359-387.

Carter, Brandon. 1974. Large Number Coincidences and the Anthropic Principle in Cosmology. *IAU Symposium 63: Confrontation of Cosmological Theories with Observational Data.* 291-298.

Carter, Brandon. 1983. The Anthropic Principle and its Implications for Biological Evolution. *Philosophical Transactions of the Royal Society of London. Series A, Mathmatical and Physical Sciences.* 310:347-363.

Dieks, D. 1992. Doomsday - Or: the Dangers of Statistics. *Philosophical Quarterly* 42(166):78-84.

Gott, R. J. 1993. Implications of the Copernican principle for our future prospects. *Nature* 363:315-319.

Leslie, J. 1989. Risking the World's End, *Bulletin of the Canadian Nuclear Society* May: 10-15.

Leslie, J. 1992. Doomsday Revisited. *Philosophical Quarterly* 42:(166):85-89.

Leslie, J. 1996. *The End of the World: The Science and Ethics of Human Extinction.* New York: Routledge.

Lewis, Antony. 2001. Comparing cosmological theories. Online. Available at http://cosmologist.info/anthropic.html.

Monton, Bradley. 2003. The Doomsday Argument Without Knowledge of Birth Rank. *Philosophical Quarterly* 53:79-82.

Nielson, H. B. 1989. Random dynamics and relations between the number of fermion generations and the fine structure constants. *Acta Physica Polonica* B20:427-468.