[A later draft of this paper is forthcoming in *Mind*. Footnotes are omitted in the HTML version.]

The Doomsday Argument is Alive and Kicking

(c) Nick Bostrom
Dept. of Philosophy, Logic and Scientific method
London School of Economics; Houghton St.; WC2A AE; London; UK
Email: n.bostrom@lse.ac.uk
Homepage: http://www.hedweb.com/nickb

Back to anthropic-principle.com

A recent paper by Korb and Oliver in this journal attempts to refute the Carter-Leslie Doomsday argument. I organize their remarks into five objections and show that they all fail. Further efforts are thus called upon to find out what, if anything, is wrong with Carter and Leslie's disturbing reasoning. While ultimately unsuccessful, Korb and Oliver's objections do however in some instances force us to become clearer about what the Doomsday argument does and doesn't imply.

Objection One

Korb and Oliver propose a minimalist constraint that any good inductive method must satisfy:

> Targeting Truth (TT) Principle: No good inductive method should—in this world—provide no more guidance to the truth than does flipping a coin. (Korb & Oliver, p. 404)

The Doomsday argument, they claim, violates this reasonable principle. In support of their claim they ask us to consider

> a population of size 1000 (i.e., a population that died out after a total of 1000 individuals) and retrospectively apply the Argument to the population when it was of size 1, 2, 3 and so on. Assuming that the Argument supports the conclusion that the total population is bounded by two times the sample value … then 499 inferences using the Doomsday Argument form are wrong and 501 inferences are right, which we submit is a lousy track record for an inductive inference schema. Hence, in a perfectly reasonable metainduction we should conclude that there is something very wrong with this form of inference. (p. 405)

But in this purported counterexample to the Doomsday argument, the TT principle is *not* violated – 501 right and 499 wrong guesses is strictly better than what one would expect to get by a random procedure such as flipping a coin. The reason why the track record is only marginally better than chance is simply that the above example assumes that the doomsayers bet on the most stringent hypothesis that they would be willing to bet on at even odds (i.e. that the total population is bounded by two times the sample value). This means, of course, that their expected utility is minimal. It is not remarkable, then, that *in this case* a person who applies the Doomsday reasoning is only slightly better off than a person who doesn't. If the bet were on the proposition not that the total population is bounded by two times the sample value but instead that it is bounded by, say, three times the sample value, then the doomsayer's advantage would be more drastic. And the doomsayer can be even more certain that the total value will not exceed thirty times the sample value.

Conclusion: Objection One does not show that the Doomsday argument violates the TT principle, nor does it show that the Doomsday reasoning at best improves the chances of being right only slightly.

Objection Two

As first noted by the French mathematician Jean-Paul Delahaye (in an unpublished manuscript), the basic Doomsday argument form can seem to be applicable not only to the survival of the human race but also to your own life span. The second of Korb and Oliver's objections picks up on this idea:

> [I]f you number the minutes of your life starting from the first minute you were aware of the applicability of the Argument to your life span to the last such minute and if you then attempt to estimate the number of the last minute using your current sample of, say, one minute, then according to the Doomsday Argument, you should expect to die before you finish reading this article. (fn. 2, p. 405)

However, this claim is incorrect. The Doomsday argument form, applied to your own life span, does not imply that you should expect to die before you've finished reading the article. The Doomsday argument says that in some cases you can reason as if you were a sample drawn randomly from a certain reference class. Taking into account the information conveyed by in this random sample, you are to update your beliefs in accordance with Bayes' theorem. This may cause a shift in your probability assignments in favor of hypotheses which imply that your position in the human race will have been fairly typical – say among the middle 98% rather than in the first or the last percentile of all humans that will ever have been born. But as John Leslie has emphasized, what probability assignment you end up with after you have made this shift, depends on your prior, i.e. the probability assignment you started out with before taking the Doomsday argument into account. In the case of the survival of the human race your prior may be based on your estimates of the risk that we will be extinguished through nuclear war, germ warfare, a disaster involving future self-replicating nanomachines, a meteor impact, etc. In the case of your own life expectancy, you will want to consider factors such as the average human life span, your state of health, and any physical danger in your environment that could cause your demise before you finish the article. Based on such considerations, the probability that you will die within the next half-hour ought presumably to strike you as extremely small. But if so, then even a considerable probability shift due to a Doomsday argument-like inference should not make you expect to die before finishing the article. Hence, contrary to what Korb and Oliver assert, the doomsayer would not make the absurd inference that she is likely to perish within an half-hour, even would she think the Doomsday argument form applicable to her individual life span.

While this is enough to refute Objection Two, the more fundamental question here is whether (and if so, how) the Doomsday argument form is applicable to individual life spans at all. I think we concede too much if we grant even a modest probability shift in this case. I have two reasons for this (which I will only outline here).

First, Korb and Oliver's application of the Doomsday argument form to individual life spans presupposes a specific solution to what has been called the problem of the reference class. Briefly stated, this is the problem of what class of entities that you should consider yourself a random sample from. Is it the class of all conscious entities? Or all entities that have a conception of their birth rank? Or all entities that are intelligent enough to be able to understand the Doomsday argument if it were explained to them? Or all entities who are in fact aware of the Doomsday argument? In my opinion, the problem of the reference class is still unsolved, and it is a serious one for the doomsayer. Korb and Oliver's objection presupposes that the reference class problem is resolved in favor of the last alternative: that the reference class consists of exactly those beings that are aware of the Doomsday argument. This might not be the most plausible solution.

The second reason for the doomsayer not to grant a probability shift in the above example is that the no-outsider requirement is not satisfied. This issue of the no-outsider requirement is somewhat complicated, but it's worth explaining because it's interesting in it's own right.

Consider first the original application of the Doomsday argument (to the survival of the human species). Suppose you were certain that there is extraterrestrial intelligent life. Let's suppose you know there is a million "small" civilizations that will have contained 200 billion persons each, and a million "large" civilizations that will have contained 200 trillion persons each. Suppose you know that the human species is one of these civilizations but you don't know whether it is small or large.

To calculate the probability that doom will strike soon (i.e. that the human species is "*Small*") we can proceed in three steps:

Step 1. Estimate the empirical prior Pr(*Small*), i.e. how likely it seems that germ warfare etc. will put an end to our species before it gets large. At this stage you don't take into account of any form of the Doomsday argument or anthropic reasoning.

Step 2. Now take account of the fact that most people find themselves in large civilizations. Let *H* be the proposition "I am a human." and define the new probability function $\Pr^*(\,.\,) = \Pr(\,.\,|\,H)$ obtained by conditionalizing on *H*. By Bayes' theorem,

$$\Pr^*(Small) = \Pr(Small|H) = \frac{\Pr(H|Small) \times \Pr(Small)}{\Pr(H)}.$$

(A similar expression holds for ¬ *Small*.) Assuming you can regard yourself a random sample from the set of all persons, we have

$$\Pr(H|Small) = \frac{200\,billion}{(200\,billion + 200\,trillion) \times 1\,million}, \text{ and}$$

$$\Pr(H|\neg Small) = \frac{200\,trillion}{(200\,billion + 200\,trillion) \times 1\,million}.$$

(If we calculate $P^*(Small)$ we find that it is very small for any reasonable prior. In other words, at this stage in the calculation, it looks as if the human species is very likely to be long-lasting.)

Step 3. Finally we take account of the Doomsday argument. Let *E* be the proposition that you find yourself "early", i.e. that you are among the first 200 billion persons in your species. Conditionalizing on this evidence, we get the posterior probability function $\Pr^{**}(\,.\,) = \Pr^*(\,.\,|\,E)$. So

$$\Pr^{**}(Small) = \Pr^*(Small|E) = \frac{\Pr^*(E|Small) \times \Pr^*(Small)}{\Pr^*(E)}$$

Note that $\Pr^*(E \mid Small) = 1$ and $\Pr^*(E \mid \neg Small) = 1/1000$. By substituting the above expressions it is then easy to verify that

$$\frac{\Pr^{**}(Small)}{\Pr^{**}(\neg Small)} = \frac{\Pr(Small)}{\Pr(\neg Small)}.$$

We thus see that we get back the empirical probabilities we started from. The Doomsday argument (in step 3) only served to cancel the effect which we took into account in step 2, namely that you were more likely to turn out to be in the human species given that the human species is one of the large rather than one of the small civilizations. This shows that if we assume that there are both "large" and "small" extraterrestrial civilizations (the precise numbers in the above example don't matter) then the right probabilities are the ones given by the naïve empirical prior. Only if we there are no "outsiders" (extraterrestrial civilizations) does the Doomsday argument work as intended.

Returning to the case where you are supposed to apply the Doomsday argument to your own life span, it appears that the no-outsider requirement is not satisfied. True, if you consider the epoch of your life during which you know about the Doomsday argument, and you partition this epoch into time-segments ("consciousness-moments"), then you might say that if you were to survive for a long time then the present moment would be an extraordinary early time-segment in this class of time-segments. You may thus be tempted to infer that you are likely to die soon (ignoring the difficulties pointed out earlier). But even if the Doomsday argument were applicable in this way, this would be the wrong conclusion to draw. For in this case you know for sure that there are many "outsiders". Here, the outsiders would be time-segments of other humans. Just as the knowledge that there are large and small extraterrestrial civilizations would annul the original Doomsday argument, so in the present case does the knowledge that there are other short-lived and long-lived humans annul the probability favoring an impending death. The fact that the present consciousness-moment belongs to you would indicate that you are an individual that will have contained many consciousness-moments rather than few, i.e. that you will be long-lived. And it can be shown (as above) that this would counterbalance the fact that the present consciousness-moment would have been extraordinarily early among all your consciousness-moments were you to be long-lived.

Conclusion: Objection Two fails to take the prior probabilities into account. These would be extremely small for the hypothesis that you will die within the next thirty minutes. Thus, contrary to what Korb and Oliver claim, even if the doomsayer thought the Doomsday argument applied to this case, he would not make the prediction that you would die within 30 minutes. However, the doomsayer should not think that the Doomsday argument is applicable in this case, for two

reasons. First, it presupposes an arguably implausible solution to the reference class problem. Second, even if we accept that only beings who know about the Doomsday argument should be in the reference class, and that it is legitimate to run the argument on time-segments of observers, the conclusion will still not follow; for the no-outsider requirement is not satisfied.

Objection Three

The third objection starts off with the claim that (in a Bayesian framework) a sample size of one is too small to make a substantial difference to one's rational beliefs.

> The main point … is quite simple: a sample size of one is "catastrophically" small. That is, whatever the sample evidence in this case may be, the prior distribution over population sizes is going to dominate the computation. The only way around this problem is to impose extreme artificial constraints on the hypothesis space. (p. 406)

They follow this claim by conceding that in a case where the hypothesis space contains only two hypotheses, a substantial shift can occur:

> If we consider the two urn case described by Bostrom, we can readily see that he is right about the probabilities. (p. 406)

(The probability in this example shifted from 50% to 99.999%, which is surely "substantial", and a similar result would be obtained for a broad range of distributions of prior probabilities.) But Korb and Oliver seem to think that such a substantial shift can only occur if we "impose extreme artificial constraints on the hypothesis space" by considering only two rival hypotheses rather than many more.

It is easy to see that this is false. Let $\{h_1, h_2, \ldots h_N\}$ be a hypothesis space and let P be any probability function that assigns a non-zero prior probability to all these hypotheses. Let h_i be the least likely of these hypotheses. Let e be the outcome of a single random sampling. Then it's easy to see, just by inspecting Bayes' formula, that the posterior probability $P(h_i \mid e)$ can be made arbitrarily big (≤ 1) by choosing e appropriately:

$$P(h_i|e) = \frac{P(e|h_i) \times P(h_i)}{\sum_{1 \leq j \leq N} \left(P(e|h_j) \times P(h_j) \right)}.$$

Choosing e such that $P(e \mid h_j)$ is small for $j \neq i$, we have

$$P(h_i|e) \approx \frac{P(e|h_i) \times P(h_i)}{P(e|h_i) \times P(h_i)} = 1$$

Indeed, we get $P(h_i \mid e) = 1$ if we choose e such that $P(e \mid h_j) = 0$ for $j \neq i$. (This would for example correspond to the case where you discover that you have a birth rank of 200 billion and immediately give probability zero to all hypotheses according to which there would be less than 200 billion persons.)

Conclusion: Korb and Oliver are wrong when they claim that the prior is always going to dominate over any computation based on a sample size of one.

Objection Four

By increasing the number of hypotheses about the ultimate size of the human species that we choose to consider, we can, according to Korb and Oliver, make the probability shift that the

Doomsday argument induces arbitrarily small:

> In any case, if an expected population size for homo sapiens ... seems uncomfortably small, we can push the size up, and so the date of our collective extermination back, to an arbitrary degree simply by considering larger hypothesis spaces. (p. 408)

The argument is that if we use a uniform prior over the chosen hypothesis space $\{h_1, h_2, ..., h_n\}$, where h_i is the hypothesis that there will have existed a total of i humans, then the expected number of humans that will have lived will depend on n: the greater the value we give to n, the greater the expected future population. Korb and Oliver compute the expected size of the human population for some different values of n and find that the result does indeed vary.

Notice first of all that nowhere in this is there a reference to the Doomsday argument. If this argument were right it would work equally against *any* way of making predictions about how long the human species will survive. For example, if during the Cuba crisis you feared – based on obvious empirical factors – that humankind might go extinct within a year, you really needn't have worried. You could just have considered a larger hypothesis space and you would thereby have reached an arbitrarily high degree of confidence that doom was not impending. If only making the world safer was that easy.

What, then, is the right prior to use for the Doomsday argument? All we can say about this from a general philosophical point of view is that it is the same as the prior for people who don't believe in the Doomsday argument. The doomsayer doesn't face a special problem here. The only legitimate way of providing the prior is through an empirical assessment of the potential threats that the human species faces. You need to base it on your best guesstimates about the hazards of germ warfare, nuclear warfare, weapons based on nanotechnology, asteroids or meteors striking the Earth, a runaway greenhouse effect, future high-energy physics experiments (which might cause a breakdown of a metastable vacuum), and other dangers as yet unimagined. (A survey of these and other risks makes up a large part of John Leslie's monograph (1996) on the Doomsday argument. He estimates the prior probability, based on these considerations, that humankind will go extinct within 200 years to be something like 5%.)

On a charitable reading, Korb and Oliver could perhaps be interpreted as saying not that Doomsday argument fails because the prior is arbitrary, but rather that the uniform prior (with some big but finite cut-off point) is as reasonable as any other prior, and that with such a prior the Doomsday argument will not show that doom is likely to strike very soon. If this is all they mean then they are not saying something that the doomsayer could not agree with. The doomsayer (i.e. a person who believes the Doomsday argument is sound) is not committed to the view that doom is likely to strike soon, only to the view that the risk that doom will strike soon is *greater* than was thought before we understood certain probabilistic implications of our having relatively low birth ranks. The Doomsday argument (if sound) shows that we have systematically underestimated the risk of doom soon, but it doesn't directly imply anything about the absolute magnitude of the probability of that hypothesis. (For example, John Leslie, who strongly believes in the Doomsday argument, still thinks there is a 70% chance that we will colonize the galaxy.) Even with a uniform prior probability, there will still be a *shift* in our probability function in favor of earlier doom.

But don't Korb and Oliver's calculations at least show that this probability shift in favor of earlier doom is in reality quite *small*, so that the Doomsday argument isn't such a big deal after all? No, their calculations do not show that, for two reasons.

The first reason is that as already mentioned, their calculations rest on the assumption of a uniform prior. Not only is this assumption gratuitous (no attempt is made to justify it) but it is also, I think, highly implausible even as an approximation of the real empirical prior. Personally I think it is fairly obvious that given what I know, the probability that there will have existed between 100 billion and 200 billion humans is much greater than the probability that there will have existed between 10^{20} and (10^{20} +100 billion) humans.

Second, even granting the uniform prior, it turns out that the probability shift is actually quite *big*. They assume a uniform distribution over the hypothesis space $\{h_1, h_2, \ldots, h_{2,048}\}$, where h_i is the hypothesis that there will have been a total of i billion humans; and they assume that you are the 60 billionth human. Then the expected size of the human population before considering the Doomsday argument is $\frac{2,048 - 60}{2} \times 10^9 =$ 994 billion. And Korb and Oliver's calculations show that after applying the Doomsday argument the expected population is 562 billion. The expected human population has been reduced by over 43% in their own example.

Conclusion: Objection Four fails. Korb and Oliver's argument about being able to get an arbitrarily large expected population by assuming a uniform prior and making the hypothesis space sufficiently big is misguided; if correct, this objection would work equally well against

predictions that do not use the Doomsday argument. For the doomsayer and the non-doomsayer use the same prior probability, the one determined by empirical considerations. Moreover, the doomsayer is not committed to the view that doom will likely strike soon, only that the risk has been systematically underestimated. Korb and Oliver have not showed that the risk has been only *slightly* underestimated. On the contrary, in Korb and Oliver's own example the Doomsday argument cuts the expected population by nearly one half.

Objection Five

Towards the end of their paper, Korb and Oliver hint at a fifth objection: that we shouldn't regard ourselves as random samples from the human species (or the human species cum its intelligent robot descendants) because there is a systematic correlation between our genetic make up and our personal identity:

> … the notion that *anyone* is uniformly randomly selected from among the total population of the species is beyond far fetched. The bodies that we are, or supervene upon, have a nearly fixed position in the evolutionary order; for example, given what we know of evolution it is silly to suppose that someone's DNA could precede that of her or his ancestors. (p. 408)

The doomsayer will grant all this. But even if the exact birth order of all humans could be inferred from a list of their genomes, the only thing that would show is that there would be more than one way of finding out about somebody's birth rank. In addition to the normal way – observing what year it is and combining that information with our knowledge of past population figures – there would now be the additional method of obtaining the same number: by analyzing somebody's DNA and consulting a table correlating DNA with birth rank.

The same holds for other correlations that may obtain. For example, the fact that I am wearing contact lenses indicates that I am living after the year 1900 A.D. This gives me a way of estimating my birth rank – check whether I have contact lenses and, if I have, draw the conclusion that it is past the year 1900 A.D.; comparing this with past population figures then tells me my birth rank. But none of these correlations add anything new once you have found at least one way of determining your birth rank.

Refusing to regard yourself as a random sample from a group just because your genes determine that you are a specific member of that group leads to implausible consequences, as the following thought experiment by John Leslie shows:

> A firm plan was formed to rear humans in two batches: the first batch to be of three humans of one sex, the second of five thousand of the other sex. The plan called for rearing the first batch in one century. Many centuries later, the five thousand humans of the other sex would be reared. Imagine that you learn you're one of the humans in question. You don't know which centuries the plan specified, but you are aware of being female. You very reasonably conclude that the large batch was to be female, almost certainly. If adopted by every human in the experiment, the policy of betting that the large batch was of the same sex as oneself would yield only three failures and five thousand successes. ... [Y]ou mustn't say: '*My genes* are female, so I have to observe myself to be female, no matter whether the female batch was to be small or large. Hence I can have no special reason for believing it was to be large.' (Leslie 1996, pp. 222-23)

If everybody were to follow the injunction that one shouldn't regard oneself as random because of one's genes, we end up with on average (5000+3) / 2 = 2,501.5 winners and equally many losers. On the doomsayer's methodology, there will be on average 5000 winners and only three losers. So at least in this case, the doomsayer's methodology is superior. This shows that, contrary to the presupposition in Objection Five, people can belong to the same reference class even if they have different genes – indeed, even if their genes differ to the point of the persons being of different genders.

(As I mentioned earlier, I think there*is* a problem for the doomsayer of exactly how and in what sense you can regard yourself as a random sample and from what class you should think

yourself as having been sampled. Leslie's thought experiment shows that it is right to regard yourself as randomly selected from some suitably defined reference class that contains people with non-identical genomes. However, many different definitions of the reference class – e.g. all persons in the experiment, all humans, all intelligent beings, all possible intelligent beings – would fit the bill for that particular thought experiment. The general problem of how to choose the reference class remains.)

Conclusion: It is true that there is a systematic correlation between our genetic makeup and our birth rank. The presence of such a correlation gives us an alternative (though impractical) way of ascertaining our birth rank but it does not affect the evidential relation between having this birth rank and any general hypothesis about humankind's future. That you can indeed legitimately regard yourself as in some sense randomly selected from a group of people even in cases where these people have different genes is shown by a thought experiment due to John Leslie (which Korb and Oliver unfortunately do not attempt to criticize or discuss). Thus, the fifth objection fails to refute the Doomsday argument.

References

Bartha, P. and Hitchcock C. 1998: "No One Knows the Date of the Hour: An Unorthodox Application of Rev. Bayes' Theorem". Paper presented at the Sixteenth Biennial Meeting of the Philosophy of Science Association.

Bostrom, N. 1997: "Investigations into the Doomsday Argument". Manuscript at http://www.anthropic-principle.com/preprints.html.

Delahaye, J-P. 1996: "Reserche de modèles pour l'argument de l'Apocalypse de Carter-Leslie". Unpublished manuscript.

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

Eckhardt, W. 1997: "A Shooting-Room view of Doomsday".*Journal of Philosophy*, Vol. XCIV, No. 5, pp. 244-259.

Kopf, T., Krtous, P, and Page, D. N. 1994: "Too Soon for Doom Gloom?".*Physics Preprints Archive*, gr-qc/9407002.

Korb, K. and Oliver, J. 1998: "A Refutation of the Doomsday Argument"*Mind*, Vol. 107, No. 426, pp. 403-410.

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