THE DOOMSDAY ARGUMENT
Nick Bostrom

Nick Bostrom presents an argument that our species will shortly disappear.

Rarely does philosophy make concrete predictions. The Doomsday argument is an important exception. From seemingly trivial premises it seeks to show that the risk that humankind will go extinct soon has been systematically underestimated. Nearly everybody's first reaction is that there must be something wrong with such an argument. Yet despite being subjected to intense scrutiny by a fair number of philosophers, no simple or trivial flaw in the argument has been identified.

Here is the doomsday argument. I will explain it in three steps:

Step I

Imagine a universe that consists of one hundred cubicles. In each cubicle, there is one person. Ninety of the cubicles are painted blue on the outside and the other ten are painted red. Each person is asked to guess whether she is in a blue or a red cubicle. (And everybody knows all this.)

Now, suppose you find yourself in one of these cubicles. What color should you think it has? Since 90% of all people are in blue cubicles, and since you don't have any other relevant information, it seems you should think that with 90% probability you are in a blue cubicle. Let's call this idea, that you should reason as if you were a random sample from the set of all observers, the *self-sampling assumption*.

Suppose everyone accepts the self-sampling assumption and everyone has to bet on whether they are in a blue or red cubicle. Then 90% of all persons will win their bets and 10% will lose. Suppose, on the other hand, that the self-sampling assumption is rejected and people think that one is no more likely to be in a blue cubicle; so they bet by flipping a coin. Then, on average, 50% of the people will win and 50% will

https://doi.org/10.1017/S1477175600002943 Published online by Cambridge University Press

lose. The rational thing to do seems to be to accept the self-sampling assumption, at least in this case.

Step II

Now we modify the thought experiment a bit. We still have the hundred cubicles but this time they are not painted blue or red. Instead they are numbered from 1 to 100. The numbers are painted on the outside. Then a fair coin is tossed (by God perhaps). If the coin falls heads, one person is created in each cubicle. If the coin falls tails, then persons are only created in cubicles 1 through 10.

You find yourself in one of the cubicles and are asked to guess whether there are ten or one hundred people. Since the number was determined by the flip of a fair coin, and since you haven't seen how the coin fell and you don't have any other relevant information, it seems you should believe with 50% probability that it fell heads (and thus that there are a hundred people).

Moreover, you can use the self-sampling assumption to assess the conditional probability of a number between 1 and 10 being painted on your cubicle given how the coin fell. For example, conditional on heads, the probability that the number on your cubicle is between 1 and 10 is 1/10, since one out of ten people will then find themselves there. Conditional on tails, the probability that you are in number 1 through 10 is one; for in that case everybody is in one of those cubicles.

Suppose that you open the door and discover that you are in cubicle number 7. Again you are asked, how did the coin fall? But now the probability is greater than 50% that it fell tails. For what you are observing is given a higher probability on the tails-hypothesis than on the heads-hypothesis. The new probability of tails can be calculated using Bayes' theorem. It is approximately 91%. So after finding that you are in cubicle number 7, you should think that with 91% probability there are only ten people.

https://doi.org/10.1017/S1477175600002943 Published online by Cambridge University Press

Step III

The last step is to transpose these results to our actual situation here on Earth. Let's formulate the following two rival hypotheses. *Doom Early*: humankind goes extinct sometime in this century and the total number of humans that will have existed is, say, 200 billion. *Doom Late*: humankind survives the century and goes on to colonize the galaxy; the total number of humans that will ever have lived is, say, 200 trillion. To simplify the exposition we will consider only these two hypotheses. (Using a more fine-grained partition of the hypothesis space wouldn't change the principle although it might modify the precise numerical values we get.)

Doom Early corresponds to there only being ten people in the thought experiment of Step II. Doom Late corresponds to there being one hundred people. Corresponding to the numbers on the cubicles, we now have the 'birth ranks' of human beings – their positions in the temporal sequence of individuals born into the human race. Corresponding to the prior probability (50%) of the coin falling heads or tails, we now have some prior probability of Doom Soon or Doom Late. This prior probability will be based on our ordinary empirical estimates of potential threats to human survival, such as nuclear or biological warfare, a meteorite destroying the planet, runaway global warming, self-replicating nanomachines running amok and so on (presumably there are dangers that we haven't yet thought of). Let's say that based on such considerations, you think that there is a 5% probability of Doom Soon. The exact number doesn't matter for the structure of the argument.

Finally, corresponding to finding you are in cubicle number 7 we have the fact that you find that your birth rank is about 60 billion (that's approximately how many humans have lived before you). Just as finding that you are in cubicle 7 increased the probability of the coin having fallen tails, so finding you are human number 60 billion gives you reason to think that Doom Soon is more probable than you previously thought. Exactly how much more probable will depend on the precise numbers you use. In the present example, the posterior prob-

https://doi.org/10.1017/S1477175600002943 Published online by Cambridge University Press

ability of Doom Soon will be very close to one. You can with near certainty rule out Doom Late.

Discussion

That is the Doomsday argument in a nutshell. After hearing about it, many people think they know what is wrong with it. But these objections tend to be mutually incompatible, and often they hinge on some simple misunderstanding. Be sure to read the literature before feeling too confident that you have a refutation.

If the Doomsday argument is correct, what precisely does it show? It doesn't show that there is no point trying to reduce threats to human survival 'because we're doomed anyway'. On the contrary, the Doomsday argument could make such efforts seem even more urgent. Taking effective action to reduce the risk that runaway global warming or nanotechnological warfare will destroy intelligent life, for example, would decrease the prior probability of Doom Soon, and this would translate into a reduction of the posterior probability of Doom Soon after the Doomsday argument is taken into account; humankind's life expectancy would go up.

There are also a number of possible 'loopholes' or alternative interpretations of what the Doomsday argument shows. For instance, if there were a dramatic decline in the world population, it would take longer before there would have been enough additional humans born to make your birth rank look surprisingly low. Alternatively, we might suppose that people in the future will be very different from the human beings of today (perhaps they will have radically transformed themselves by the use of some advanced technology); and in that case one might argue that these hypothetical future beings should be in a different reference class from ourselves. A more careful formulation of the self-sampling assumption would state that you should reason as if you were a random sample from the set of all observers *in your reference class*. If the reference class were suitably restricted, then the Doomsday argument would be stopped in its tracks.

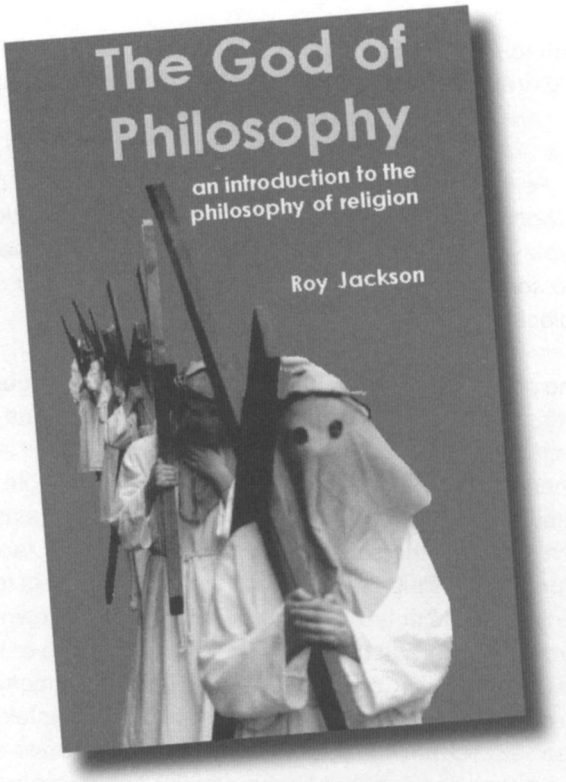
https://doi.org/10.1017/S1477175600002943 Published online by Cambridge University Press

My own view is that the Doomsday argument is inconclusive, and the problem has to do with determination of the reference class. But it is by no means a trivial matter. We cannot just arbitrarily pick any reference class we like. It is possible, using a variety of arguments and thought experiments, to establish some fairly strong constraints on what would constitute a reasonable choice of reference class. For example, if you decided to use an extremely narrow reference class, one that included only observers that are in exactly the same subjective epistemic situation as you, then you would make it impossible to derive any observational predictions at all from many contemporary cosmological theories, despite the fact that cosmologists appear to be testing these theories against experimental data all the time.

Exactly how to handle the choice of reference class remains an important open research question. In my book *Anthropic Bias*, I argued that this question needs to be addressed as part of a wider effort to develop a general theory of observation selection effects. Such a theory would explain how we should reason when either the evidence we have, or the hypothesis that we are trying to evaluate, contains indexical information of the sort that is central to the Doomsday argument – information about *who* you are, *where* you are, or *what time* it is now. A probabilistic theory of reasoning about such information would not only help us assess the Doomsday argument, it would also have applications in many other philosophical and scientific areas: the Sleeping Beauty problem, game theory problems involving imperfect recall, the foundations of quantum physics, cosmological theories, and even some questions in evolutionary biology, just to mention a few. A better understanding of observational selection effects will impose surprising constraints on beliefs about the future of our species and about the distribution of observers in the universe.

Dr Nick Bostrom is in the faculty of philosophy and is director of the Future of Humanity Institute at Oxford University. His research interests are wide-ranging, and preprints of many of his publications can be found at www.nickbostrom.com.

https://doi.org/10.1017/S1477175600002943 Published online by Cambridge University Press



The God of Philosophy is a comprehensive, engaging and accessible introduction to the philosophy of religion. Covering the concept of God, arguments for and against the existence of God, faith and reason, religious language and life after death, the book is set to become the first word for students and general readers alike.

'...a timely addition to resources ... Illuminating without being overbearing.'
Times Educational Supplement

	To order, phone	
In the UK	**(01442) 879097**	**£9.99**
In North America	**1-800-444-2419**	**$15.99**

or visit www.philosophers.co.uk/tpmshop.htm

https://doi.org/10.1017/S1477175600002943 Published online by Cambridge University Press