Causality and the Doomsday Argument

Ivan Phillips
The Futurists Guild

January 24th 2005

Abstract

Using the Autodialer thought experiment, we show that the Self-Sampling Assumption (SSA) is too general, and propose a revision to the assumption that limits its applicability to causally-independent observers. Under the revised assumption, the Doomsday Argument fails, and the paradoxes associated with the standard SSA are dispelled. We also consider the effects of the revised sampling assumption on tests of cosmological theories. There we find that, while we must restrict our attention to universes containing at least one observer, the total number of observers predicted in each universe is irrelevant to the confirmation of a theory.

The Doomsday Argument (Bostrom 1997, 2002; Leslie 1989, 1996) concludes that the prior probability we assign to the short-term extinction of humanity is greatly magnified by our observation that we exist at the present time in human history. Before we discuss the Doomsday Argument itself, it is instructive to examine one of the thought experiments that is held up as its analogue. The following experiment, Cubicles, is a minor variation of the original Incubator described by Bostrom (2002).

> Cubicles
>
> One hundred cubicles are numbered 1 to 100. A coin is flipped to determine how many people will be placed in the cubicles. If the result is heads, one person is placed in all 100 cubicles. If the result is tails, only the first 10 cubicles are occupied. Each person who enters a cubicle is led in blindfolded so they cannot tell which cubicle they are placed in, nor can they tell how many other people have been placed in cubicles. Initially, you find yourself in a cubicle, but you do not know which one. You are asked to estimate the probability that the coin toss was tails. You have no additional information, so you answer 50%.
>
> Next, you learn that you are in cubicle #5. What is the posterior probability that the coin flip was tails?

The revised probability of the coin flip having been tails can be computed using Bayes' theorem:

$$p(\text{tails}|\#5) = p(\#5|\text{tails})\, p(\text{tails}) \,/\, (p(\#5|\text{tails})\, p(\text{tails}) + p(\#5|\text{heads})\, p(\text{heads})) \qquad (1)$$

$$= 10\%\ 50\% \,/\, (10\%\ 50\% + 1\%\ 50\%)$$

$$= 10 / 11 = 91\%$$

This calculation is a natural outcome of the fact that occupancy in cubicle #5 is ten times more likely if the coin flip was tails rather than heads.

To get to the Doomsday Argument, one must argue that the Cubicles scenario is analogous to the question of how many humans will ever be born. Suppose we assign every human individual a unique birth rank. Adam and Eve would be numbered one and two, respectively, and contemporary humans would have birth ranks in the 60 billion range (to date, about 60 billion humans have ever lived). To construct the Doomsday Argument, we suppose that a finite number of humans will ever live, and theorize that humanity will either

go extinct sometime in the next century (Doom Soon), or else humans will go on to colonize the galaxy and go extinct after thousands of years (Doom Late). Let's say that Doom Soon is the equivalent of the last human to be born having a birth rank of 100 billion, and that Doom Late equates to the last birth having rank of 100 trillion.

Suppose that our birth ranks are analogous to the cubicle numbers in the Cubicles experiment. The Doomsday Argument is the corresponding claim that we should consider our own observation of our birth rank (about 60 billion) as being a single, random sample of all possible observations of one's birth rank. If the claim is true, then, as in the Cubicles experiment, Bayes' theorem tells us that the prior probability of Doom Soon is greatly amplified. Unless we had prior reason to believe that Doom Late was a thousand times more improbable than Doom Soon, we should now expect Doom Soon to be more likely.

The Doomsday Argument is cited as an exemplary application of the Self-Sampling Assumption (Bostrom 2000):

> (SSA) Observers should reason as if they were a random sample from the set of all observers in their reference class.

In the case of the Cubicles experiment, the observers (the cubicle occupants) certainly consider themselves as being part of a random sample of all possible occupants. But can we say the same of humans in the Doomsday Argument?

The most compelling criticism of the Doomsday Argument was put forward by Sowers (2002). Sowers was able to show that the sampling of birth ranks in the Doomsday Argument is not random, and Sowers explicitly cites the time-sequencing of observations as the mechanism responsible for robbing the Doomsday Argument of its random sampling.

In answer to Sowers, Bostrom (2002B) responds with some thought experiments that show that temporally-sequencing a set of observations isn't sufficient to destroy the statistical validity of its sampling. Instead, having previously shown that paradoxes can arise as a result of the SSA (Bostrom 2001), Bostrom proposes a revised version of the SSA called the *Strong Self-Sampling Assumption*:

> (SSSA) 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.

Bostrom goes on to argue that the SSSA renders the Doomsday Argument inconclusive. Unfortunately, the SSSA also has the side effect of removing the probability shift in the Cubicles experiment, contradicting equation (1).

A complete understanding of these matters must resolve several issues. First, we must determine precisely what it is about the sequencing of observations that makes the Doomsday Argument invalid. Second, we must identify what restrictions we must place on the Self-Sampling Assumption to make its use consistent with results from probability theory (e.g., with the Cubicles experiment). In no case should we find the restricted SSA to arrive at results that contradict standard probability theory. Finally, we must determine whether those same restrictions eliminate the paradoxes cited by Bostrom.

The Autodialer Experiments

To make the heterology between the Cubicles experiment and the Doomsday Argument clear, let's consider a new thought experiment: the Autodialer. The Autodialer experiment has several variations which show us when and how the SSA can be applied.

Experiment 1: Intelligent Autodialer

At some location remote from you, there are 100 phone booths numbered from 1 to 100. Secretly, a coin has been flipped. If it landed tails, then 10 people have been placed into the first 10 booths. If the coin landed heads, then all of the booths are populated. The coin is fair, so your prior probability that the coin landed heads is 50%. You have an autodialer that will dial, at random, one of the *populated* phone booths. You make a phone call and the occupant of the dialed phone booth tells you he is in phone booth #5. What is the posterior probability that the coin landed heads?

This experiment is isomorphic to the Cubicles experiment, and, the posterior probability calculation is identical to equation (1):

$$p(\text{tails}|\text{dial \#5}) = p(\text{dial \#5}|\text{tails})\, p(\text{tails}) \,/\, (p(\text{dial \#5}|\text{tails})\, p(\text{tails}) + p(\text{dial \#5}|\text{heads})\, p(\text{heads}))$$

$$= 10\%\ 50\% \,/\, (10\%\ 50\% + 1\%\ 50\%)$$

$$= 10/11 \qquad (2)$$

There's a 91% probability that the coin flip was tails given that someone in booth #5 answered the random call to a populated booth.

Let's consider what would happen if we replaced our Intelligent Autodialer with a Random Autodialer.

Experiment 2: Random Autodialer

The Random Autodialer dials phone booths at random, whether or not they are occupied. If the phone booth is unoccupied, an answering machine responds with a recorded message and the number of the booth that was dialed. You place a call and the occupant of booth #5 answers. What is the probability that the coin landed tails?

The Random Autodialer will dial phone booth #5 no more frequently than any other phone booth, no matter what the outcome of the coin toss. That is $p(\#5|\text{tails}) = p(\#5|\text{heads})$. Since both coin flips will place a person in phone booth #5, our posterior probability is the same as our prior probability:

$$p(\text{tails}|\text{dial \#5}) = p(\text{dial \#5}|\text{tails})\, p(\text{tails}) \,/\, (p(\text{dial \#5}|\text{tails})\, p(\text{tails}) + p(\text{dial \#5}|\text{heads})\, p(\text{heads}))$$

$$= 1\%\ 50\% \,/\, (1\%\ 50\% + 1\%\ 50\%)$$

$$= 1/2 = p(\text{tails}) \qquad (3)$$

Thus, the Random Autodialer cannot tell anything from the fact that it called one of the first 10 booths. The two theories, heads and tails, predict different distributions of occupants, not different distributions of phone booths. However, the Random Autodialer is taking a random sample of phone booths, which both theories predict to be uniformly distributed from 1 to 100.

Finally, we implement a Sequential Autodialer.

Experiment 3: Sequential Autodialer

The Sequential Autodialer dials only the occupied booths in sequence from 1 to 10 (or 1 to 100).

You place the first call and the occupant of booth #1 answers. What is the probability that the coin landed tails?

Case 3B: You make 5 calls. On the fifth call, the occupant of booth #5 answers. What is the probability that the coin landed tails?

We can immediately see that dialing booth #1 on the first call doesn't tell us anything about the number of booths that are occupied because p(dial #1|tails) = p(dial #1|heads) = 100%.

$$p(\text{tails}|\text{dial \#1}) = p(\text{dial \#1}|\text{tails})\, p(\text{tails}) / (p(\text{dial \#1}|\text{tails})\, p(\text{tails}) + p(\text{dial \#1}|\text{heads})\, p(\text{heads}))$$

$$= 100\%\ 50\% /\ (100\%\ 50\% + 100\%\ 50\%)$$

$$= 1/2 = p(\text{tails}) \qquad (4)$$

Likewise, the probability of dialing booth #2 after having previously dialed booth #1 is also unity. Indeed, we don't learn anything about the coin toss until we make the eleventh phone call, despite the fact that we are only calling populated booths. From a probability theoretical point of view, the Sequential Autodialer fails to gain information from the first 10 samples because it is not *randomly* sampling from the population probability distributions of the two theories.

Only the Intelligent Autodialer is sensitive to the occupancy of all of the phone booths, so it is the only variant that can be used to learn something about the coin flip without sampling booths 11 to 100. This particular autodialer plays the role of the existence of the observer in the Cubicles experiment. Just as the Intelligent Autodialer cannot dial an empty phone booth, so the cubicle occupant cannot find herself in booth #5 and also find that booth #5 is empty.

We can see that the scenario depicted in the Doomsday Argument is homologous to the Sequential Autodialer, not the Intelligent Autodialer. We cannot know our own birth rank before all of those humans before us have sampled their own birth ranks in sequential order. Even if we initially have no idea of our own birth rank, our rank can only be measured (or provided to us) by having counted-off birth rank observations from the first human. Unlike the numbers painted on the outside of the cubicles, our birth rank observations are causally connected with those of all previous observers. By analogy, we see that the Doomsday Argument fails for the reason cited by Sowers: its sampling is not random. The posterior probability of Doom Soon is exactly equal to the prior probability of Doom Soon.

Of course, we could combine elements of the Intelligent Autodialer and the Doomsday Argument to form the Time-Traveling Phone Booth experiment:

Time-Traveling Phone Booth

Your Time-Traveling Phone Booth travels across space and time such that you will meet a person with a random birth rank. When you pick up the receiver, you are transported to room where a man is seated. The man tells you that his birth rank is 90 billion. Should you expect Doom Soon?

The answer is yes, Doom Soon is much more likely. Though the man in the room had no particular expectation for Doom Soon based on his birth rank, you, the time traveler now have a random birth rank sample to amend your probabilities.

Multiple Occupancy

The aforementioned thought experiments can teach us even more when we start adding multiple occupants to our cubicles and phone booths. We can ask what would happen to the Intelligent Autodialer experiment if we placed 10 occupants in each of the first 10 phone booths, no matter what the result of the coin flip. During the first phone call, we might speak to any number of occupants of booth #5, and each of them will reply that they are in booth #5. The probability calculations remain unchanged despite the fact that we have

varied the total number of observers predicted by each theory.

The analogous change to the cubicles experiment would be to place two occupants in each of the first 10 cubicles regardless of the coin toss. As long as we have a mechanism to ensure that, before the cubicle number is revealed, the occupants cannot tell how many people share their cubicle, the prior probabilities remain the same. Let's say that the occupants are blindfolded until it comes time to reveal their cubicle number. Again, this modification does not alter the probability calculation.

In contrast to this simple result, a literal reading of the SSA would have us change our posterior probability assessment based on the number of blindfolded occupants who share our cubicle. So it seems that the original SSA is inconsistent with the results we would expect in the case of multiple occupancy.

Implications for the SSA

The Autodialer and Doomsday thought experiments appear to place limits on where we can apply the SSA.

First, we must be certain that our observations are randomly sampling the probability distributions which we are trying to test. The Doomsday Argument failed on this account because it treats birth ranks as if they were random samples when they are not random.

Second, we cannot double-count observers who are constrained by physics to simultaneously make exactly the same observations.

We propose incorporating these conditions of causal independence into the SSA, creating the Causally-Independent Self-Sampling Assumption:

> (CISSA) Causally-distinct observers should reason as if they were a random sample from the set of all causally-distinct observers in their reference class.

Causality in this context is a constraint imposed on the observations of one observer by the observations of another observer. When physics forces the observations to occur in a non-random sequence or renders the observers indistinguishable to the theories under test, then the observers are not causally-distinct.

The demand that observers be grouped with their causally-indistinct peers has the effect of eliminating the paradoxes associated with the standard version of the SSA. One such paradox is the Adam and Eve paradox, quoted here from Bostrom (2001):

> *Eve and Adam, the first two humans, knew that if they gratified their flesh, Eve might bear a child, and if she did, they would be expelled from Eden and would go on to spawn billions of progeny that would cover the Earth with misery. One day a serpent approached the couple and spoke thus: "Pssst! If you embrace each other, then either Eve will have a child or she won't. If she has a child then you will have been among the first two out of billions of people. Your conditional probability of having such early positions in the human species given this hypothesis is extremely small. If, one the other hand, Eve doesn't become pregnant then the conditional probability, given this, of you being among the first two humans is equal to one. By Bayes' theorem, the risk that she will have a child is less than one in a billion. Go forth, indulge, and worry not about the consequences!"*

The flaw in the serpent's argument is that it ignores the physical connection between Adam and Eve and their potential descendants. Not only are Adam and Eve's descendants constrained to measure their birth ranks in order, they are also constrained to succeed the birth ranks of their parents. This causal connection violates the CISSA assumption, and the serpent's argument fails. Similarly, the UN++ and Quantum Joe paradoxes (Bostrom 2001) are dispelled under the CISSA.

Consequences for Cosmological Tests

Each cosmological theory predicts a probability distribution for finding some universe as a function of the physical constants of that universe. Every life form that evolves in a given universe must be consistent with the physics of that universe, and every intelligent observer in that universe is constrained to observe the same physical constants. Hence, CISSA tells us that each observer in a given universe is constrained to observe the same thing and should count as just a single observer.

Suppose we have two cosmological theories, A and B which predict different probability distributions for creating a universes with different physical constants. Let's denote the physical constants of a universe by θ. Some of these universes will be habitable by intelligent species like us. Let's consider the case where the odds of life forming in a universe with a given set of physical constants is dependent only on those physical constants, and not the specific cosmological theory that predicted those constants.

We do not have a good estimate the odds of intelligent life L arising for a given θ, that is, we do not know $p(L|\theta)$. However, it will turn out to be unimportant, as long as we restrict our attention to universes to those where it is non-zero. For each model, we can assume that the probability that we would find intelligent life in a universe with a given θ is the product of the probability that life will arise given θ and the probability that a universe with that θ will be found at all in that model:

$$p(L \text{ in } \theta|A) = p(L|\theta)\, p(\theta|A) \tag{5}$$

Using Bayes' theorem, we can calculate the likelihood of theory A given the fact that intelligent life exists in a universe with constants θ:

$$p(A|L \text{ in } \theta) = p(L \text{ in } \theta|A)\, p(A)/ (p(L \text{ in } \theta|A)\, p(A) + p(L \text{ in } \theta|B)\, p(B)) \tag{6}$$

$$= p(L \mid \theta)\, p(\theta|A)\, p(A)/ (p(L \text{ in } \theta)\, p(\theta|A)\, p(A) + p(L|\theta)\, p(\theta|B)\, p(B))$$

$$= p(\theta|A)\, p(A)/ (p(\theta|A)\, p(A) + p(\theta|B)\, p(B))$$

That is, $p(L|\theta)$ drops out of the calculation. Therefore, we find that:

$$p(A|L \text{ in } \theta) = p(A|\theta) \tag{7}$$

This is exactly what we would expect from our initial assumption that the possibility of life is fixed on θ. In essence, our assumption is that there is a physical constant that represents the compatibility of life with all the other parameters. This "Life Compatibility Quotient" is a function only of the other observed physical parameters.

Our resulting expression for our posterior probability $p(A|L \text{ in } \theta)$ depends only on the ratio of $p(\theta_0|A)$ and $p(\theta_0|B)$ where θ_0 represents the physical constants of our own universe. However, we must be sure to normalize our prior probabilities correctly. In (4), the prior probabilities $p(\theta|A)$ and $p(\theta|B)$ are given by probability distributions normalized over all universes compatible with intelligent life, i.e., with $p(L|\theta) > 0$.

$$\int_{p(L|\theta) > 0} p(\theta|A)\, d\theta = 1 \tag{8}$$

$$\int_{p(L|\theta) > 0} p(\theta|B)\, d\theta = 1$$

Due to this normalization, we should prefer theories that favor θ_0 over theories that, say, maximize the total number of observers over all θ with $p(L|\theta) > 0$. For example, suppose $p(\theta|B)$ is nonzero for all inhabitable universes whereas $p(\theta|A)$ is nonzero only for θ_0. In that case, the observational data confirms theory A, not theory B, even though the total number of observers in B will be larger. This result is in contrast with what we would expect under the standard SSA. The SSA weights the prior probabilities by $p(L|\theta)$ so that theories which favor highly populated universes are preferred over theories that favor universes physically like our own.

Is the assumption of equation (5) a valid one? It is claimed here that this assumption is always valid. It is difficult to see how two scientific theories could predict different values for $p(L|\theta)$. Surely, θ would parameterize any possible physical or historical factors that would impact the evolution of life. If we discovered some new kind of physical phenomena that influenced the likelihood of life evolving, we would have to either explain it as a function of our existing parameters or invent a new physical constant. For example, if we determined that the universe was subject to some previously unknown cosmological contraction that limited the amount of time that life had to evolve, we would most certainly parameterize the effect. The observation that we exist is not causally-independent of the observation of the physical parameters of our universe.

Thus, if we accept the CISSA, our posterior probabilities do not depend on the number of intelligent observers in each universe.

Conclusion

We derive a a new Causally-Independent Self-Sampling Assumption (CISSA) by restricting the SSA to groups of observers who can make random, causally-independent observations. The CISSA is consistent with physics, and ensures that we don't double-count observers. Under the CISSA, the Doomsday Argument fails, and the paradoxes associated with the standard SSA are dispelled.

Consequently, it is shown that the posterior probabilities we compute for cosmological models are independent of the number of observers in each universe. However, we must be sure to correctly normalize our prior probability distributions over all universes which have at least one observer.

References

Bostrom, N. (1997) *Investigations into the Doomsday argument, Preprint* at http://www.anthropicprinciple.com/preprints/inv/investigations.html

Bostrom, N. (2000) *Observational Selection Effects and Probability*. PhD thesis, London School of Economics, 2000.

Bostrom, N. (2001). *The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe,* Preprint, Synthese , Vol. 127, Issue 3.

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

Bostrom, N. (2002B) *Beyond the Doomsday Argument: Reply to Sowers and Further Remarks*, http://www.anthropic-principle.com/preprints/sowers/beyond.pdf

Franceschi, Paul (2003) *A Third Route to the Doomsday Argument*, http://cogprints.org/2990/

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

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

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