Observation Selection Theory and Cosmological Fine-tuning

Nick Bostrom
Faculty of Philosophy, Oxford University
www.nickbostrom.com

1. Introduction

When our measurement instruments sample from only a subspace of the domain that we are seeking to understand, or when they sample with uneven sampling density from the target domain, the resulting data will be affected by a selection effect. If we ignore such selection effects, our conclusions may suffer from selection biases. A classic example of this kind of bias is the election poll taken by the Literary Digest in 1936. On the basis of a large survey, the *Digest* predicted that Alf Langdon, the Republican presidential candidate, would win by a large margin. But the actual election resulted in a landslide for the incumbent, Franklin D. Roosevelt. How could such a large sample size produce such a wayward prediction? The *Digest*, it turned out, had harvested the addresses for its survey mainly from telephone books and motor vehicle registries. This introduced a strong selection bias. The poor of the depression era – a group that disproportionally supported Roosevelt – often did not have phones or cars.

The Literary Digest suffered a major reputation loss and soon went out of business. It was superseded by a new generation of pollsters, including George Gallup, who not only got the 1936 election right but also managed to predict what the *Digest's* prediction would be to within 1%, using a sample size that was only one thousandth as large. The key to his success lay in his accounting for known selection effects. Statistical techniques are now routinely used to correct for many kinds of selection bias.

Observation selection effects are an especially subtle kind of selection effect that is introduced not by limitations in our measurement apparatuses but by the fact that all evidence is preconditioned on the existence of an observer to "have" the evidence and to build the instruments in the first place. Only quite recently have observation selection effects become the subject of systematic study. Observation selection effects are important in many scientific areas, including cosmology and parts of evolution theory, thermodynamics, the foundations of quantum theory, and traffic analysis. There are also interesting applications to the search for extraterrestrial life and questions such as whether we might be living in a computer simulation created by an advanced civilization [1].

Observation selection theory owes a large debt to Brandon Carter, a theoretical physicist who wrote several seminal papers on the subject, the first one published in 1974 [2-5]. Although there were many precursors, one could fairly characterize Carter as the father of observation selection theory – or "anthropic reasoning" as the field is also known. Carter coined the "weak" and the "strong anthropic principle", intending them to express injunctions to take observation selection effects into account. But while Carter knew how to apply his principles to good effect, his explanations of the methodology they were meant to embody were less than perfectly clear. The meaning of the anthropic principles was further obscured by some later interpreters, who bestowed them with additional content that had nothing to do with observation selection effects. This contraband content, which was often of a speculative, metaphysical, or teleological nature, caused "anthropic" reasoning to fall into disrepute. Only recently has this trend been reversed.

Since Carter's contributions, considerable effort has been put into the working out of the applications of anthropic principles, especially as they pertain to cosmological fine-tuning. There have also been many philosophical investigations into the foundations of anthropic reasoning. These investigations have revealed several serious paradoxes, such as the Doomsday argument (which one may or may not regard as paradoxical) [6], the Sleeping Beauty problem [7] [8], and the Adam and Eve thought experiments [9]. It is still controversial what conclusions we should draw from the apparent fine-tuning of our universe, as well as whether and to what extent our universe really is fine-tuned, and even what it means to say that it is fine-tuned.

Developing a theory of observation selection effects that caters to legitimate scientific needs while sidestepping philosophical paradoxes is a non-trivial challenge. In my recent book *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, I presented the first mathematically explicit general observation selection theory and explored some of its implications.

Before sketching some of the basic elements of this theory and discussing how it pertains to the multiverse hypothesis, let us briefly consider some of the difficulties that confront attempts to create a method for dealing with observation selection effects.

2. The need for a probabilistic anthropic principle

The anthropic principles that Carter proposed, even setting aside the inadequacies in their formulation, were insufficiently strong for many scientific applications. A particularly serious shortcoming is that they were not probabilistic.

Carter's principles enable us to deal with some straightforward cases. Consider a simple theory that says that there are 100 universes, and that 90 of these are lifeless and 10 contain observers. What does such a theory predict that we should observe? Clearly not a lifeless universe. Since lifeless universes contain no observers, an observation selection effect, enunciated by the strong anthropic principle, precludes them from being observed. So although the theory claims that the majority of universes are lifeless, it nevertheless predicts that we should observe one of the atypical ones that contain observers.

Now take a slightly more complicated case. Suppose a theory says that there are 100 universes of the following description:

90 type-A universes, which are lifeless
9 type-B universes, which contain one million observers each
1 type-C universe, which contains one billion observers

What does this theory predict that we should observe? (We need to know the answer to this question in order to determine whether it is confirmed or disconfirmed by our observations.) As before, an obvious observation selection effect precludes type-A universes from being observed, so the theory does not predict that we should observe one of those. But what about type-B and type-C universes? It is logically compatible with the theory that we should be observing a universe of either of these kinds. However, probabilistically it is more likely, conditional on the theory, that we should observe the type-C universe, because that's what the theory says that over 99% of all observers observe. Finding yourself in a type-C universe would in many cases tend to confirm such a theory, to at least some degree, compared with other theories that imply that most observers live in type-A or type-B universes.

To get this result, we must introduce a probabilistic strengthening of the anthropic principle along the lines of what I have called the *Self-Sampling Assumption* [8, 10, 11]:

> (SSA) One should reason as if one were a random sample from the set of all observers in one's reference class.[1]

With the help of SSA, we can calculate the conditional probabilities of us making a particular observation given one theory or another, by comparing what fraction of the observers in our reference class would be making such observations according to the competing theories.

What SSA does is enable us to take indexical information into account. Consider the following two evidence statements concerning the cosmic microwave background radiation (CMB):

> *E*: An observation of CMB = 2.7K is made.
> *E**: *We* make an observation of CMB = 2.7K.

Note that *E** implies *E*, but not *vice versa*. *E**, which includes a piece of indexical information, is logically stronger than *E*. It is consequently *E** that dictates what we should believe in case these different evidence statements lead to different conclusions. This follows from the principle that all relevant information should be taken into account.

Let us examine a case where it is necessary to use *E** rather than *E* [18]. Consider two rival theories about the local temperature of CMB. Let T_1 be the theory we actually hold, claiming that CMB = 2.7K. Let T_2 say that CMB = 3.1K. Now, suppose that the universe is infinitely large and contains an infinite number of stochastic processes of suitable kinds, such as radiating black holes. If for each such random process there is a finite, non-zero probability that it will produce an observer in any particular brain state (subjectively making an observation *e*), then, because there are infinitely many independent "trials", the probability, for any given observation *e*, that *e* will be made by some observer somewhere in the universe is equal to 1. Let *B* be the proposition that this is the case. We might wonder how we could possibly test a conjunction like $T_1\&B$, or $T_2\&B$. For whatever observation *e* we make, both these conjunctions predict equally well (with probability 1) that *e* should be made. According to Bayes's theorem, this entails that conditionalizing on *e* being made will not affect the probability of $T_1\&B$, or of $T_2\&B$. And yet it is obvious that the observations we have actually made support $T_1\&B$ over

[1] Related principles have also been explored in e.g. [12-15]; see also [16, 17]

$T_2\&B$, for, needless to say, it is because of our observations that we believe that CMB = 2.7K and not 3.1K.

This problem is solved by going to the stronger evidence statement E^* and applying SSA. For any reasonable choice of reference class, $T_1\&B$ implies that a much larger fraction of all observers in that class should observe CMB = 2.7K than CMB = 3.1K, than does $T_2\&B$. (According to $T_1\&B$, all normal observers observe CMB = 2.7K, while on $T_2\&B$ only some exceptional black-hole-emitted observers, or those who suffer from rare illusions, observe CMB = 2.7K.) Given these facts, SSA implies:

$$P(E^* \mid T_1\&B) >> P(E^* \mid T_2\&B) \qquad (1)$$

From (1) it is then easy to show that our actual evidence E^* does indeed give us reason to believe in $T_1\&B$ rather than $T_2\&B$. In other words, SSA makes it possible for us to know that CMB = 2.7.

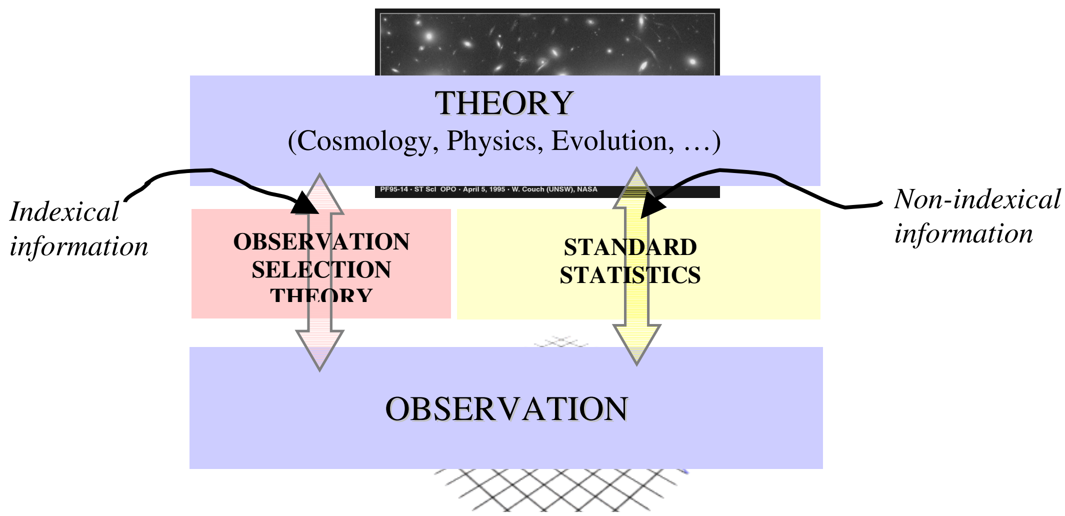

Figure 1. Observation selection theory is a complement to standard statistics, needed to handle cases where either the evidence or the hypothesis includes indexical information.

For the moment we are setting aside the problem of exactly how the reference class is to be defined. In the above example, any reference class definition satisfying some very weak constraints would do the trick. To keep things simple, we also ignore the problem of how to generalize SSA to deal with infinite domains. Strictly speaking, such an extension, which might involve focusing on densities rather than sets of observers, would be necessary to handle the present example; but it would add complications that would distract from basic principles.

3. Challenges for observation selection theory

So far, so good. SSA can derive additional support from various thought experiments, and it can be applied to a number of scientific problems where it yields results that are less obvious but nevertheless valid.

Unfortunately, if we use SSA with the universal reference class, the one consisting of all intelligent observers, we encounter paradoxes. One of these is the notorious Doomsday argument, which purports to show that we have systematically underestimated the probability that our species will go extinct soon. The basic idea behind this argument is that our position in the sequence of all humans that will ever have lived (roughly number 60 billion) would be much more probable if the total number of humans is, say, 200 billion rather than 200 trillion. Once we take into account this difference in the conditional probability of our observed birth rank, the argument goes, hypotheses that imply that very many humans are yet to be born are seen to be much less probable than we would have thought if we only considered the ordinary evidence (about the risk of germ warfare, nuclear was, meteor strikes, destructive nanotechnology etc.) The prospects of our descendants ever colonizing the galaxy would be truly dismal, as this would make our own place in human history radically atypical.

The most common initial reaction to the Doomsday argument is that it must be wrong; moreover, that it is wrong for some obvious reason. Yet when it comes to explaining *why* it is wrong, it turns out that there are almost as many explanations as there are people who disbelieve the Doomsday arguments. And the explanations tend to be mutually inconsistent. On closer inspection, all these objections, which allege some trivial fallacy, turn out to be themselves mistaken [6, 8, 19].

Nevertheless, the Doomsday argument has some backers, and while the way in which it aims to derive its conclusion is definitely counterintuitive, it may not quite qualify as a paradox. It is therefore useful to consider the following thought experiment [9], which has a structure similar to the Doomsday argument but yields a conclusion that is even harder to accept.

> *Serpent's Advice.* Eve and Adam, the first two humans, knew that if they gratified their flesh, Eve might bear a child, and that if she did, they would both be expelled from Eden and go on to spawn billions of progeny that would fill the Earth with misery. One day a serpent approached the couple and spoke thus: "Pssssst! If you take each other in carnal embrace, then either Eve will have a child or she won't. If she has a child, 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 does *not* become pregnant then the conditional probability, given this, of you being among the first two humans is equal to one. By Bayes's theorem, the risk that she shall bear a child is less than one in a billion. Therefore, my dear friends, step to it and worry not about the consequences!"

It is easy to verify that, if we apply SSA to the universal reference class, the serpent's mathematics is watertight. Yet surely it would be irrational for Eve to conclude that the risk of her becoming pregnant is negligible.

One can try to revise SSA in various ways or to impose stringent conditions on its applicability. However, it is difficult to find a principle that satisfies all constraints that an observation selection theory ought to satisfy – a principle that both serves legitimate scientific needs and at the same time is probabilistically coherent and paradox-free. Here we lack the space to elaborate on the multitude of such theory constraints. It is easy enough to formulate a theory that passes a few of these tests but it is hard to find one that survives them all.

4. Sketch of a solution

The solution, in my view, begins with the realization that the problem with SSA is not that it is too strong but that it isn't strong enough. SSA tells you to take into account one kind of indexical information: information about which observer you are. But you have more indexical information than that. You also know *which temporal segment* of that observer, which "observer-moment", you currently are. We can formulate a *Strong Self-Sampling Assumption* that takes this information into account [8]:

> (SSSA) Each observer-moment should reason as if it were randomly selected from the class of all observer-moments in its reference class.

Arguments can be given for why SSSA expresses a correct way of reasoning about a number of cases.

To cut a long story short, we find that the added analytical firepower provided by SSSA makes it possible to relativize the reference class, so that different observer-moments of the same observer may place themselves in different reference classes without that observer being probabilistically incoherent over time. This relativization of the reference class in turn makes it possible coherently to reject the serpent's advice to Eve while still enabling legitimate scientific inferences to go through. Recall, for instance, the case we considered above, about our observation of CMB = 2.7K supporting the theory that this is the actual local temperature of CMB even when evaluated in the context of a cosmological theory that asserts that all possible human observations are made. This result would be obtained almost independently of how we defined the reference class. So long as the reference class satisfies some very weak constraints, the inference works. This "robustness" of an inference under different definitions of the reference class turns out to be a hallmark of those applications of anthropic reasoning that are scientifically respectable. By contrast, the applications that yield paradoxes rely on specific definitions of the reference class and collapses when a different reference class chosen. The serpent's reasoning, for example, works only if we place the observer-moments of Adam and Eve prior to sinning in the same reference class as the observer-moments of those (very different) observers that may come to exist centuries later as a result of the first couple's moral lapse. The very fact that this absurd consequence would follow from selecting such a reference class gives us a good reason to use another reference class instead.

The idea expressed vaguely in SSSA can be formalized into a precise principle that specifies the evidential bearing of a body of evidence *e* on a hypothesis *h*. I have dubbed this the *Observation Equation* [8]:

$$P_\alpha(h \mid e) = \frac{1}{\gamma} \sum_{\sigma \in \Omega_h \cap \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_\alpha \cap \Omega(w_\sigma)|}$$

Here, α is the observer-moment whose subjective probability function is P_α. Ω_h is the class of all possible observer-moments about whom *h* is true; Ω_e is the class of all possible observer-moments about whom *e* is true; Ω_α is the class of all observer-moments that α places in the same reference class as herself; w_α is the possible world in which α is located; and γ is a normalization constant. The quantity in the denominator is the cardinality of the intersection of two classes, Ω_α and $\Omega(w_\sigma)$, the latter being the class of all observer-moments that exist in the possible world w_σ.

The Observation Equation can be generalized to allow for different observer-moments within the reference class having different weights $\mu(\sigma)$. This option is of particular relevance in the context of the many-worlds version of quantum mechanics, where the weight of an observer-moment would be proportional to the amplitude squared of the branch of the universal wavefunction where that observer-moment lives.

The Observation Equation expresses the core of a quite general methodological principle. Two of its features deserve to be highlighted here. The first is that by dividing the terms of the sum by the denominator, we are factoring out the fact that some possible worlds contain more observer-moments than do others. If one omitted this operation, one would in effect assign a higher prior probability to possible worlds that contain a greater number of observers (or more long-lived observers). This would be equivalent to accepting the *Self-Indication Assumption*, which prescribes an a priori bias towards worlds that have a greater population. But although the Self-Indication Assumption has its defenders (e.g. [20]), it leads to paradoxical consequences, as shown by the Presumptuous Philosopher thought experiment [8]. In particular, it implies that we should assign probability 1 to the cosmos being infinite, even if we had strong empirical evidence that it was finite; and this implication is very hard to accept.

A second feature to highlight is that the only possible observer-moments that are taken into account by an agent are those that the agent places in the same reference class as itself. Observer-moments that are outside this reference class are treated, in a certain sense, as if they were rocks or other lifeless objects. Thus, the question of how to define "observer" is replaced with the question of how an agent should select an appropriate reference class for a particular application. This reference class will often be a proper subset of intelligent observers or observer-moments.

Bounds can be established on permissible definitions of the reference class. For example, if we reject the serpent's advice, we must not use the universal reference class that places all observer-moments in the same reference class. If we want to conclude on the basis of our evidence that CMB = 2.7K, we must not use the minimal reference class that includes only subjectively indistinguishable observer-moments, for such a reference class would block that inference.

It is an open question whether additional constraints can be found that would always guarantee the selection of a unique reference class for all observer-moments or whether there might instead, as seems quite likely, be an unavoidable element of subjective judgment in the choice of reference class. This latter contingency would parallel the widely acknowledged element of subjectivity inherent in many other kinds of scientific judgments that are made on the basis of limited or ambiguous evidence.

5. Implications for cosmological fine-tuning

One immediate implication of observation selection theory for cosmological fine-tuning is that it allays worries that anthropic reasoning is fundamentally unsound and inevitably plagued by paradoxes. It thereby puts the multiverse explanation of fine-tuning on a more secure methodological footing.

A multiverse theory can potentially explain cosmological fine-tuning, provided several conditions are met. To begin with, the theory must assert the existence of an ensemble of physically real universes. The universes in this ensemble would have to differ from one another with respect to the values of the fine-tuned parameters, according to a suitably broad distribution. If observers can exist only in those universes in which the relevant parameters take on the observed fine-tuned values (or if the theory at least implies that a large portion of all observers are likely to live in such universes), then an observation selection effect can be invoked to explain why we observe a fine-tuned universe. Moreover, in order for the explanation to be completely satisfactory, this postulated multiverse should not itself be significantly fine-tuned. Otherwise the explanatory problem would merely have been postponed; for we would then have to ask, how come the *multiverse* is fine-tuned? A multiverse theory meeting these conditions could give a relatively high conditional probability to our observing a fine-tuned universe. It would thereby gain a measure of evidential support from the finding that our universe is fine-tuned. Such a theory could also help *explain* why we find ourselves in a fine-tuned universe, but to do in this, the theory would also have to meet the ordinary crew of desiderata – it would have to be physically plausible, fit the evidence, be relatively simple and non-gerrymandered, and so forth. Determining whether this potential anthropic explanation of fine-tuning actually succeeds requires a lot of detailed work in empirical cosmology.

One may wonder whether these conclusions depend on fine-tuning *per se* or whether they follow directly from the generic methodological injunction that we should, other things being equal, prefer simpler theories with fewer free variables to more complex theories that require a larger number of independent stipulations to fix their parameters (Occam's razor). In other words, how does the fact that *life would not have existed if the constants of our universe had been slightly different* play a role in making fine-tuning cry out for an explanation and in suggesting a multiverse theory as the remedy?

Observation selection theory helps us answer these questions. It is not just that all single-universe theories in the offing would seem to require delicate handpicking of lots of independent variable-values that would make such theories unsatisfactory: the fact that life would not otherwise have existed adds to the support for a multiverse theory. And how does that fact do this? By making the anthropic multiverse explanation possible. A simple multiverse theory could potentially give a high conditional probability to us observing the kind of universe we do because it says that only that kind of universe, among all the universes in a multiverse, would be observed (or at least, that it would be observed by a disproportionately large fraction of the observers). The observation selection effect operating on the fact of fine-tuning *concentrates* the conditional probability on us observing a universe like the actual one.

Further, observation selection theory enables us to answer the question of how big a multiverse has to be in order to explain our evidence. The upshot is that bigger is not always better [8]. The postulated multiverse would have to be large and varied enough to make it probable that some universe like ours should exist. Once this objective is reached, there is no additional anthropic ground for thinking that a theory that postulates an even bigger ensemble of universes is therefore, other things equal, more probable. The choice between two multiverse theories that both give a high probability to a fine-tuned universe like ours existing must be made on other grounds, such as simplicity or how well they fit with the rest of physics.

A multiverse would not have to be large enough to make it probable that a universe *exactly* like ours should exist. A multiverse theory that entails such a huge cosmos that one would expect a universe exactly like ours to be included in it, does not have an automatic advantage over a more frugal competitor. Such an advantage would have to be earned, for example by being a simpler theory. There is, as we noted earlier, no general reason for assigning a higher probability to theories that entail that there is a greater number of observers in our reference class. Increasing the membership in our reference class might make it more likely that the reference class should contain some observer who is making exactly the observations that we are making but it would also make it more surprising that we should happen to be that particular observer rather than one of the others in the reference class. The net effect of these two considerations is to cancel each other out. All the observation selection effect does is concentrate conditional probability on the observations represented by the observer-moments in our reference class so that, metaphorically speaking, we can postulate stuff outside the reference class "for free". Postulating additional stuff *within* the reference class is not gratis in the same way but would have to be justified on independent grounds.

It is, consequently, in major part an empirical question whether a multiverse theory is more likely than a single-universe theory, and whether a larger multiverse is more plausible than a smaller one. Anthropic considerations are an essential part of the methodology for addressing these questions, but the answers will depend on the data.

In its current stage of development, observation selection theory falls silent on problems where the solution depends sensitively on the choice of reference class. For example, suppose a theory implies that the overwhelming majority of all observers that exist are of a very different kind from us. Should these radically different observers be in our reference class? If we do place them in our reference class (or more precisely, if we place their observer-moments in the same reference class as our own current observer-moments), then a theory that implies that the overwhelming majority of all observers are of that different kind would be contraindicated by our evidence, roughly because according to that theory we should have thought it highly unlikely that we should have found ourselves to be the kind of observer that we are rather than a more typical kind of observer. That is to say, such a theory would be disconfirmed compared to an equally simple theory that implied that a much larger fraction of all observers would be of our kind. Yet if we exclude the other kind of observer from our reference class, then our evidence would not count against the theory. In a case like this, the choice of reference class makes a difference to our interpretation of our evidence.

Further developments of observation selection theory would be needed to determine whether there is a unique objectively correct way of resolving such cases. In the meantime, it is a virtue of the methodological framework encapsulated by the Observation Equation that it brings this indeterminacy into the open and does not surreptitiously privilege one particular reference class over potentially equally defensible alternatives.

References

Preprints and background material on observation selection theory can be found at ***www.anthropic-principle.com***

1. Bostrom, N., "Are You Living in a Computer Simulation?," *Philosophical Quarterly* 53(211) (2003): pp. 243-55.
2. Carter, B., *Large Number Coincidences and the Anthropic Principle in Cosmology*, in *Confrontation of Cosmological Theories with Data*, M.S. Longair, Editor. 1974, Reidel: Dordrecht. pp. 291-298.
3. Carter, B., "The Anthropic Principle and its Implications for Biological Evolution," *Philosophical Transactions of the Royal Society* A 310 (1983): pp. 347-363.
4. Carter, B., *The Anthropic Selection Principle and the Ultra-Darwinian Synthesis*, in *The Anthropic Principle*, F. Bertola and U. Curi, Editors. 1989, Cambridge University Press: Cambridge. pp. 33-63.
5. Carter, B., *Large Number Coincidences and the Anthropic Principle in Cosmology*, in *Physical Cosmology and Philosophy*, J. Leslie, Editor. 1990, Macmillan Publishing Company.
6. Bostrom, N., "The Doomsday Argument is Alive and Kicking," *Mind* 108(431) (1999): pp. 539-550.
7. Bostrom, N., "Sleeping Beauty: A Synthesis of Views," *working paper* (2004)
8. Bostrom, N., *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (New York: Routledge, 2002).
9. Bostrom, N., "The Doomsday argument, Adam & Eve, UN++, and Quantum Joe," *Synthese* 127(3) (2001): pp. 359-387.
10. Bostrom, N., "Investigations into the Doomsday argument," *Preprint* http://www.anthropic-principles.com/preprints/inv/investigations.html (1997)
11. Bostrom, N., "Observer-relative chances in anthropic reasoning?," *Erkenntnis* 52 (2000): pp. 93-108.
12. Gott, R.J., "Implications of the Copernican principle for our future prospects," *Nature* 363(27 May) (1993): pp. 315-319.
13. Vilenkin, A., "Predictions From Quantum Cosmology," *Physical Review Letters* 74(6) (1995): pp. 846-849.
14. Page, D.N., "Sensible Quantum Mechanics: Are Probabilities Only in the Mind?," *International Journal of Modern Physics D* 5(6) (1996): pp. 583-596.
15. Page, D.N., *Can Quantum Cosmology Give Observational Consequences of Many-Worlds Quantum Theory*, in *General Relativity and Relativistic Astrophysics, Eighth Canadian Conference, Montreal, Quebec*, C.P. Burgess and R.C. Myers, Editors. 1999, American Institute of Physics: Melville, New York. pp. 225-232.
16. Tegmark, M., "Is 'the theory of everything' merely the ultimate ensemble theory?," *Physics preprints archive* gr-gc/9704009(3 Apr) (1997)

17. Linde, A. and A. Mezhlumian, "On Regularization Scheme Dependence of Predictions in Inflationary Cosmology," *Phys.Rev. D* 53 (1996): pp. 4267-4274.
18. Bostrom, N., "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation," *Journal of Philosophy* 99(12) (2002): pp. 607–623.
19. Leslie, J., *The End of the World: The Science and Ethics of Human Extinction* (London: Routledge, 1996).
20. Olum, K., "The Doomsday Argument and the Number of Possible Observers," *Philosophical Quarterly* 52(1) (2002): pp. 164-84.