

# The Mysteries of Self-Locating Belief and Anthropic Reasoning

By Nick Bostrom

§1. HOW BIG IS THE SMALLEST FISH IN THE POND? YOU TAKE YOUR WIDE-MESHED fishing net and catch one hundred fish, every one of which is greater than six inches long. Does this evidence support the hypothesis that no fish in the pond is much less than six inches long? Not if your wide-meshed net can't actually catch smaller fish.

The limitations of your data collection process affect the inferences you can draw from the data. In the case of the fish-size-estimation problem, a *selection effect*—the net's being able to sample only the big fish—invalidates any attempt to extrapolate from the catch to the population remaining in the water. Had your net had a finer mesh, allowing it to sample randomly from *all* the fish, then finding a hundred fish all greater than a foot long would have been good evidence that few if any fish remaining were much smaller.

In the fish net example, a selection effect is introduced by the fact that the instrument you used to collect data sampled from only a subset of the target population. Analogously, there are selection effects that arise not from the limitations of the measuring device but from the fact that all observations require the existence of an appropriately positioned observer. These are known as *observation selection effects*.

The study of observation selection effects is a relatively new discipline. In my recent book *Anthropic Bias*, I have attempted to develop the first mathematically explicit theory of observation selection effects. In this article, I will attempt to convey a flavor of some of the mysteries that such a theory must resolve.

The theory of observation selection effects may have implications for a number of fields in both philosophy and science. One example is evolutionary biology, where observation selection effects must be taken into ac-

*Nick Bostrom is a British Academy Postdoctoral Fellow and a Research Fellow in Philosophy at Oxford University. His research interests include philosophy of science, foundations of probability theory, and ethics of technology and science. His first book, Anthropic Bias: Observation Selection Effects in Science and Philosophy, came out in 2002. His other recent publications include "Self-Locating Belief in Big Worlds" (Journal of Philosophy, 2002) and "Are You Living In a Computer Simulation?" (Philosophical Quarterly, 2003).*

count when addressing questions such as the probability of intelligent life developing on any given Earth-like planet. We know that intelligent life evolved on Earth. Naively, one might think that this piece of evidence suggests that life is likely to evolve on most Earth-like planets, but that would overlook an observation selection effect. No matter how small the proportion of all Earth-like planets that evolve intelligent life, *we* must be from a planet that did (or we must be able to trace our origin to a planet that did, if we were born in a space colony) in order to be an observer ourselves.

Our evidence—that intelligent life arose on our planet—is therefore predicted equally well by the hypothesis that intelligent life is very improbable even on Earth-like planets as it is by the hypothesis that intelligent life is highly probable on Earth-like planets. The evidence does not distinguish between the two hypotheses, provided that in both hypotheses intelligent life would very likely have evolved somewhere.

**§2.** ANOTHER EXAMPLE COMES FROM COSMOLOGY, WHERE OBSERVATION SELECTION effects are crucial considerations in deriving empirical predictions from the currently popular so-called ‘multiverse theories’, according to which our universe is but one out of a vast ensemble of physically real universes out there.

Some cases are relatively straightforward. 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? Obviously not that we should observe a lifeless universe. Because lifeless universes contain no observers, an observation selection effect precludes them from being observed. So although the theory says that the majority of universes are lifeless, it nevertheless predicts that we should observe one of the atypical ones that contain observers.

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

90 type-A universes: they are lifeless.

9 type-B universes: they contain one million observers each.

1 type-C universe: it contains one billion observers.

What does this theory predict that we should observe? (We need to know that 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 99% of all observers observe.

Couldn’t we hold instead that the theory predicts that we should observe a type-B universe? After all, it says that type-B universes are much more common than those of type-C. There are various arguments that show

that this line of reasoning is untenable. We lack the space to review them all here, but we can hint at one of the underlying intuitions by considering an analogy. Suppose you wake up after having been sedated and find yourself blindfolded and with earplugs. Let's say for some reason you come to consider two rival hypotheses about your location: that you are somewhere on the landmass of Earth, or that you are at sea. You have no evidence in particular to suggest that you should be at sea, but you are aware that there are more square meters of sea than of land. Clearly, this does not give you ground for thinking you are at sea. For you know that the vast majority of observers are on land, and in the absence of more specific relevant evidence to the contrary, you should think that you probably are where the overwhelming majority of people like you are.

In a similar vein, the cosmological theory that says that almost all people are in type-C universes predicts that you should find yourself in such a universe. Finding yourself in a type-C universe would in many cases tend to confirm such a theory, to at least some degree, compared to other theories that imply that most observers live in type-A or type-B universes.

**§3.** LET US NOW LOOK A LITTLE MORE SYSTEMATICALLY AT THE REASONING ALLUDED to in the foregoing paragraphs. Consider the following thought experiment:

*Dungeon:* The world consists of a dungeon that has one hundred cells. In each cell there is one prisoner. Ninety of the cells are painted blue on the outside and the other ten are painted red. Each prisoner is asked to guess whether he is in a blue or a red cell. (And everybody knows all this.) You find yourself in one of these cells. What color should you think it is?

*Answer: Blue, with 90% probability.*

Since 90% of all observers are in blue cells, and you don't have any other relevant information, it seems that you should set your credence (that is, your subjective probability, or your degree of belief) of being in a blue cell to 90%. Most people seem to agree that this is the correct answer. Since the example does not depend on the exact numbers involved, we have the more general principle that in cases like this, your credence of having property *P* should be equal to the fraction of observers who have *P*. You reason *as if* you were a randomly selected observer. This principle is known as the *Self-Sampling Assumption (SSA)*:

SSA: One should reason as if one were a random sample from the set of all observers in one's reference class.<sup>1</sup>

For the time being, we can assume that the reference class consists of all intelligent observers, although this is an assumption that needs to be revised, as we shall see later.

While many accept without further argument that *SSA* is applicable to *Dungeon*, let's briefly consider how one might seek to defend this view if challenged to do so. One argument one can adduce is the following: suppose

that everyone accepts *SSA* and everyone has to bet on whether they are in a blue or a red cell. Then 90% of the prisoners will win their bets; only 10% will lose. If, on the other hand, *SSA* is rejected, and the prisoners think that one is no more likely to be in a blue cell than in a red cell, and they bet, for example, by flipping a coin, then on average merely 50% of them will win and 50% will lose. It seems better that *SSA* be accepted.

What allows the people in *Dungeon* to do better than chance is that they have a relevant piece of empirical information regarding the distribution of observers over the two types of cells; they have been informed that 90% are in blue cells. It would be irrational not to take this information into account. We can imagine a series of thought experiments where an increasingly large fraction of observers are in blue cells: 91%, 92%, ..., 99%. As the situation gradually degenerates into the limiting 100% case where they are simply told, "You are all in blue cells," from which each prisoner can deductively infer that he is in a blue cell, it is plausible to require that the strength of prisoners' beliefs about being in a blue cell should gradually approach probability one. *SSA* has this property.

These considerations support the initial intuition about *Dungeon*: that it is a situation in which one should reason in accordance with *SSA*.

One thing worth noting about *Dungeon* is that we didn't specify how the prisoners arrived in their cells. The prisoners' history is irrelevant so long as they don't know anything about it that gives them clues about the color of their cells. For example, they may have been allocated to their respective cells by some objectively random process such as by drawing balls from an urn (while blindfolded so they couldn't see where they ended up). Or they may have been allowed to choose cells for themselves, a fortune wheel subsequently being spun to determine which cells should be painted blue and which red. But the thought experiment doesn't depend on there being a well-defined randomization mechanism. One may just as well imagine that prisoners have been in their cells since the time of their birth, or indeed since the beginning of the universe. If there is a possible world in which the laws of nature determine, without any appeal to initial conditions, which individuals are to appear in which cells and how each cell is painted, then the inmates would still be rational to follow *SSA*, provided only that they did not have knowledge of the laws or were incapable of deducing what the laws implied about their own situation. Objective chance, therefore, is not an essential ingredient of the thought experiment; it runs on low-octane subjective uncertainty.

**§4.** SO FAR, SO GOOD. IN *DUNGEON*, THE NUMBER OF OBSERVERS FEATURING IN THE experiment was fixed. Now let us consider a variation where the total number of observers depends on which hypothesis is true. This is where the waters begin to get treacherous.

*Incubator. Stage (a):* The world consists of a dungeon with one hundred cells. The cells are numbered on the outside consecutively from 1 to 100. The numbers cannot be seen from inside the cells. There is also a mechanism called

“the incubator.” The incubator first creates one observer in cell #1. It then flips a coin. If the coin lands tails, the incubator does nothing more. If the coin lands heads, the incubator creates one observer in each of the remaining ninety-nine cells as well. It is now a time well after the coin was tossed, and everyone knows all the above. Stage (b): A little later, you are allowed to see the number on your cell door, and you find that you are in cell #1.

*Question: What credence should you give to tails at stages (a) and (b)?*

We shall consider three different models for how to reason, each giving a different answer. These three models may *appear* to exhaust the range of plausible solutions, although we shall later outline a fourth model which is the one that in fact I think points to the way forward.

*Model 1.* At stage (a) you should set your credence of tails equal to 50%, since you know that the coin toss was fair. Now consider the *conditional* credence you should assign at stage (a) to being in a certain cell given a certain outcome of the coin toss. For example, the conditional probability of being in cell #1 given tails is 1, since that is the only cell you can be in if that happened. And by applying SSA to this situation, we get that the conditional probability of being in cell #1 given heads is 1/100. Plugging these values into the well-known mathematical result known as Bayes’s theorem, we get:

$$\begin{aligned} &Pr(\text{tails} \mid \text{I am in cell \#1}) \\ &= \frac{Pr(\text{I am in cell \#1} \mid \text{tails})Pr(\text{tails})}{Pr(\text{I am in cell \#1} \mid \text{tails})Pr(\text{tails}) + Pr(\text{I am in cell \#1} \mid \text{heads})Pr(\text{heads})} \\ &= \frac{1 \times 1/2}{1 \times 1/2 + 1/100 \times 1/2} = 100/101 \end{aligned}$$

Therefore, upon learning that you are in cell #1, you should become almost certain (Pr = 100/101) that the coin fell tails.

*Answer: At stage (a) your credence of tails should be 1/2 and at stage (b) it should be 100/101.*

Now consider a second model that sort of reasons in the opposite direction:

*Model 2:* Since you know the coin toss to have been fair, and you haven’t got any other relevant information, your credence of tails at stage (b) should be 1/2. Since we know the conditional credences (same as in model 1), we can infer, via Bayes’s theorem, what your credence of tails should be at stage (a), and the result is that your prior credence of tails must equal 1/101.

*Answer: At stage (a) your credence of tails should be 1/101 and at stage (b) it should be 1/2.*

Finally, we can consider a model that denies that you gain any relevant information from finding that you are in cell #1:

*Model 3: Neither at stage (a) nor at stage (b) do you have any relevant information as to how the coin fell. Thus in both instances, your credence of tails should be 1/2.*

*Answer: At stage (a) your credence of tails should be 1/2 and at stage (b) it should be 1/2.*

**§5.** LET US TAKE A CRITICAL LOOK AT THESE THREE MODELS. WE SHALL BE EGALITARIAN and present one problem for each of them.

We begin with model 3. The challenge for this model is that it seems to suffer from incoherency. It is easy to see (simply by inspecting Bayes's theorem) that if we want to end up with the posterior probability of tails being 1/2, and both heads and tails have a 50% prior probability, then the conditional probability of being in cell #1 must be the same on tails as it is on heads. But at stage (a), you know with certainty that if the coin fell heads then you are in cell #1, so this conditional probability must equal 1. In order for model 3 to be coherent, you would therefore have to set your conditional probability of being in cell #1 given heads equal to 1 as well. That means you would already know with certainty at stage (a) that you are in cell #1, which is simply not the case! Hence, we must reject model 3.

Readers who are familiar with David Lewis's 'Principal Principle'<sup>2</sup> may wonder if it is not the case that model 3 is firmly based on this principle, so that rejecting model 3 would mean rejecting the 'principal principle' as well. That is not so. While this is not the place to delve into the details of the debates about the connection between objective chance and rational credence, suffice it to say that the 'principal principle' does not state that you should always set your credence equal to the corresponding objective chance if you know it. Instead, it says that you should do this *unless* you have other relevant information that needs to be taken into account.<sup>3</sup> There is some controversy about how to specify which sorts of such additional information will modify reasonable credence when the objective chance is known, and which sorts of additional information leaves the identity intact. But there is wide agreement that the proviso is needed. Now, in *Incubator* you do have such extra relevant information that you need to take into account, and model 3 fails to do that. The extra information is that at stage (b), you have discovered that you were in cell #1. This information is relevant because it bears probabilistically on whether the coin fell heads or tails; or so, at least, the above argument seems to show.

**§6.** MODEL 1 AND MODEL 2 ARE BOTH ALL RIGHT AS FAR AS PROBABILISTIC COHERENCE goes. Choosing between them would therefore be a matter of selecting

the most plausible or intuitively appealing prior credence function.

Model 2 says that at stage (a) you should assign a credence of  $1/101$  to the coin having landed tails. That is, just knowing about the setup but having no direct evidence about the outcome of the toss, you should be virtually certain that the coin fell in such a way as to create 99 additional observers. This amounts to having an a priori bias towards the world containing many observers. Modifying the thought experiment by using different numbers, it can be shown that in order for the probabilities always to work out the way model 2 requires, you would have to subscribe to the principle that, other things being equal, a hypothesis that implies that there are  $2N$  observers should be assigned twice the credence of a hypothesis that implies that there are only  $N$  observers. This principle is known as the *Self-Indication Assumption (SIA)*.<sup>4</sup> My view is that this assumption is untenable. To see why, consider the following example (which seems to be closely analogous to *Incubator*):

*The Presumptuous Philosopher:* It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories:  $T_1$  and  $T_2$  (using considerations from super-duper symmetry). According to  $T_1$ , the world is very, very big but finite and there are a total of a trillion trillion observers in the cosmos. According to  $T_2$ , the world is very, very, *very* big but finite and there are a trillion trillion trillion observers. The super-duper symmetry considerations are indifferent between these two theories. Physicists are preparing a simple experiment that will falsify one of the theories. Enter the presumptuous philosopher: "Hey guys, it is completely unnecessary for you to do the experiment, because I can already show you that  $T_2$  is about a trillion times more likely to be true than  $T_1$ !" (whereupon the philosopher explains model 2 and appeals to *SIA*).

Somehow one suspects that the Nobel Prize committee would be reluctant to award the philosopher the big one for this contribution. Yet it is hard to see what the relevant difference is between this case and *Incubator*. If there is no relevant difference, and we are not prepared to accept the argument of the presumptuous philosopher, then we are not justified in using model 2 in *Incubator* either.

**§7.** WHAT ABOUT MODEL 1, THEN? IN THIS MODEL, AFTER FINDING THAT YOU ARE in cell #1, you should set your credence of tails equal to  $100/101$ . In other words, you should be almost certain that the world does not contain the extra 99 observers. This might seem like the least unacceptable of the alternatives, and therefore the one we ought to go for. However, before we uncork the bottle of champagne, ponder what this option entails:

*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 them and spoke thus: "Pssst! If you hold each other, 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, on 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, indulge your desires and worry not about the consequences!"

Given the assumption that the same method of reasoning should be applied as in *Incubator*, and using some plausible prior probability of pregnancy given carnal embrace (say, approximately 1/100), it is easy to verify that there is nothing wrong with the serpent's mathematics. The question, of course, is whether the assumption should be granted.

Let us review some of the differences between *Incubator* and *Serpent's Advice* to see if any of them are relevant in the sense of providing a rational ground for treating the two cases differently:

*i. In the Incubator experiment there was a point in time, stage (a), when the subject was actually ignorant about her position among the observers. By contrast, Eve presumably knew all along that she was the first woman.*

But it is not clear why that should matter. We can imagine that Eve and Adam were created on a remote island, and that they didn't know whether there are other people on Earth, until one day they were informed that they are thus far the only ones. It is still counterintuitive to say that the couple needn't worry about the possibility of Eve getting pregnant.

*ii. When the subject is making the inference in Incubator, the coin has already been tossed. In the case of Eve, the relevant chance event has not yet taken place.*

This difference does not seem crucial either. We can modify *Serpent's Advice* by supposing that the deciding chance event has already taken place. Let's say the couple has just sinned and they are now brooding over the ramifications. Should the serpent's argument completely reassure them that nothing bad will happen? It seems not. So the worry remains.

*iii. At stage (b) in Incubator, any observers resulting from the toss have already been created, whereas Eve's potential progeny do not yet exist at the time when she is assessing the odds.*

We can consider a variant of *Incubator* where each cell exists in a different century. That is, let us suppose that cell #1, along with its observer, are cre-

ated in the first century, and destroyed after, say, 30 years. In each of the subsequent 99 centuries, a new cell is built, allowed to exist for 30 years, and is then destroyed. At some point in the first century a coin is tossed and, depending on how it lands, these subsequent cells will or will not contain observers. Stage (*a*) can now be defined to take place in the first century after the first prisoner has been created but before the coin has been tossed and before the prisoner has been allowed to come out of his cell to observe its number. At this stage (stage (*a*)), it seems that he should assign the same credence to tails and the same conditional credences of tails given that he is in a particular cell as he did in the original version—for precisely the same reasons. But then it follows, just as before, that his posterior credence of tails, after finding that he is in cell #1, should be much greater than the prior credence of tails. This version of *Incubator* is analogous to *Serpent's Advice* with respect to the non-existence of the later humans at the time when the odds are being assessed.

*iv. In Incubator, the two hypotheses under consideration (heads and tails) have well-defined known prior probabilities (50%), whereas Eve and Adam must rely on vague subjective considerations to assess the risk of pregnancy.*

True, but would we want to say that if Eve's getting pregnant were determined by some distinct microbiological process with a well-defined objective chance which Eve and Adam knew about, then they ought to accept the serpent's advice? If anything, the knowledge of such an objective chance would make the consequence even weirder.

**§8.** THE MYSTERY THAT WE ARE FACING HERE IS THAT IT SEEMS CLEAR THAT BOTH the serpent and the presumptuous philosopher are wrong, yet it seems as if the only model that yields this double result (model 1) is incoherent. One may be tempted to blame the strength of *SSA* for these troubles and think that we should reject it. But that, it appears, would transfix us on another horn of the dilemma, for we would then have to reject the cogent argument about the *Dungeon* thought experiment presented above, and, perhaps even more seriously, we would have failed to account for a number of very well-founded scientific applications in cosmology and elsewhere (which I lack the space to fully explore in this article).

There are a number of possible moves and objections that one can try at this point. But most of these maneuvers and objections rest on simple misunderstandings, or else they fail to provide a workable alternative to how to reason about the range of problems that need to be addressed. It is easy enough to come up with a method of reasoning that works in one particular case, but when one then tests it against other cases—philosophical thought experiments and legitimate scientific inferences—one usually soon discovers that it yields paradoxes or otherwise unacceptable results. Yet by seriously confronting this central conundrum of self-locating belief, we can glean important clues about what a general theory of observation selection effects must look like.

§9. SO WHERE DO WE GO FROM HERE? THE FULL ANSWER IS COMPLICATED AND difficult and cannot be fully explored in a relatively short paper like this one. But by helping myself to a fair amount of hand-waving, I can at least try to indicate the direction in which I think the solution is to be found.

One key to the solution is to realize 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 a certain kind of indexical information—information about which observer you are. But you have more indexical information than that about who you are: you also know *when* you are. That is, you know which temporal segment—which “observer-moment”—of an observer that you are at the current time. We can formulate a ‘Strong Self-Sampling Assumption’ that takes this information into account:

*SSSA*: Each observer-moment should reason as if it were randomly sampled from its reference class.

Arguments can be given for *SSSA* along lines parallel to those of the arguments for *SSA* provided above. For example, one can consider cases in which a person is unaware of what time it is and has to assign credence to different temporal possibilities.

A second key to the solution is to see how the added analytical power of *SSSA* enables us to relativize the reference class. What this means is that different observer-moments of the same observer may use different reference classes without that observer being incoherent over time. To illustrate, let us again consider the *Incubator* thought experiment. Before, we rejected model 3 because it seemed to imply that the reasoner should be incoherent. But we can now construct a new model, model 4, which agrees with the answers that model 3 gave, that is, a credence of 1/2 of heads at both stage (*a*) and stage (*b*), but which modifies the reasoning that led to these answers in a such a way as to avoid incoherency.

Suppose that just as before and for the same reasons, we assign, at stage (*a*), the credences:

$$\begin{aligned} \Pr(\text{tails}) &= 1/2 \\ \Pr(I'm \text{ in cell } \#1 \mid \text{tails}) &= 1 \\ \Pr(I'm \text{ in cell } \#1 \mid \text{heads}) &= 1/100 \end{aligned}$$

Now, if the *only* epistemic difference between stage (*a*) and stage (*b*) is that at the latter stage you have the additional piece of information that you are in cell #1, then Bayesian conditionalization of the above conditional credences entails (as in model 1) that your posterior credence must be:

$$\Pr_{\text{posterior}}(\text{tails}) = \Pr(\text{tails} \mid I'm \text{ in cell } \#1) = 100/101$$

However, when we take *SSSA* into account, we see that there are other epistemic differences between stages (*a*) and (*b*). In addition to gaining the information that you are in cell #1, you also *lose* information when you enter stage (*b*). At stage (*a*), you knew that you were currently an observer-mo-

ment who is ignorant about which cell you are in and who is pondering different possibilities. At stage (b), you no longer know this piece of indexical information, because it is no longer true of you that you currently are such an observer-moment. You do know that you are an observer who *previously* was at stage (a), but this is an indexically different piece of knowledge from knowing that you are currently at stage (a). Since your total information at stage (b) is not equal to the information you had at stage (a) conjoined with the proposition that you are in cell #1, there is therefore no requirement that your beliefs at stage (b) be obtained by conditionalizing your stage (a) credence function on the proposition that you are in cell #1.

Normally, this kind of subtle change in indexical information makes no difference to our inferences, so they can therefore usually be ignored. In special cases, however, including the thought experiments considered in this paper, which rely precisely on the peculiar evidential properties of indexical information, such changes can be highly relevant.

This does not yet show that your beliefs at stage (b) about the outcome of the coin toss should differ from those obtained by conditionalizing  $\Pr(\text{tails} | \text{I'm in cell \#1})$ , but it defeats the Bayesian argument for why they should be the same. If you regard these associated epistemic changes that occur in addition to your obtaining the information that "I'm in cell #1" when you move from stage (a) to stage (b) as relevant, then you can coherently assign a 1/2 posterior credence to tails.

Let  $\hat{a}$  be one of your observer-moments that exist before you discover which cell you are in. Let  $\hat{a}$  be one of your observer-moments that exist after you have discovered that you are in cell #1 (but before you have learned about the outcome of the coin toss). What probabilities  $\hat{a}$  and  $\hat{a}$  assign to various hypotheses depends on reference classes in which they place themselves. For example,  $\hat{a}$  can pick a reference class consisting of the observer-moments who are ignorant about which cell they are in, while  $\hat{a}$  can pick the reference class consisting of all observer-moments who know they are in cell #1.  $\hat{a}$ 's conditional credences are then the same as before:

$$\begin{aligned} \Pr_{\hat{a}}(\hat{a} \text{ is in cell \#1} \mid \text{tails}) &= 1 \\ \Pr_{\hat{a}}(\hat{a} \text{ is in cell \#1} \mid \text{heads}) &= 1/100 \end{aligned}$$

But  $\hat{a}$ 's conditional probability of being in cell #1 given heads is now identical to that given tails:

$$\begin{aligned} \Pr_{\hat{a}}(\hat{a} \text{ is in cell \#1} \mid \text{tails}) &= 1 \\ \Pr_{\hat{a}}(\hat{a} \text{ is in cell \#1} \mid \text{heads}) &= 1 \end{aligned}$$

From this, it follows that  $\hat{a}$ 's posterior credence of tails after conditionalizing on  $\hat{a}$  being in cell #1 is the same as its posterior credence of heads, namely 1/2.

SSSA does not by itself *imply* that this should be  $\hat{a}$ 's posterior credence of tails. It just shows that it is a coherent position to take. The actual credence assignment depends on which reference classes are chosen. In the

case of *Incubator*, it may not be obvious which choice of reference class is best. But in the *Serpent's Advice*, it is clear that Eve should select a reference class that puts her observer-moments existing at the time when she is pondering the possible consequences of the sinful act in a different reference class from those later observer-moments that may come to exist as a result of her transgression. For her to do otherwise would not be incoherent, but it would yield the strongly counterintuitive consequence discussed above. By selecting the more limited reference class, she can reject this consequence.

The question arises whether it is possible to find some general principle that determines what reference class an observer-moment should use. We may note that the early Eve's choice of a reference class that contains only her own early observer-moments and excludes the observer-moments of all the billions of progeny that may come to exist later is not completely arbitrary. After all, the epistemic situation that the early Eve is in is very different from the epistemic situation of these later observer-moments. Eve doesn't know whether she will get pregnant and whether all these other people will come to exist; her progeny, by contrast, would have no doubts about these issues. Eve is confronted with a very different epistemic problem than her possible children would be. It is thus quite natural to place Eve in a different reference class from these later people, even apart from the fact that this maneuver would explain why the serpent's recommendation should be eschewed.

Constraints on what could be legitimate choices of reference class can be established, but it is an open question whether these will always suffice to single out a uniquely correct reference class for every observer-moment. My suspicion is that there might remain a subjective element in the choice of reference class in some applications. Furthermore, I suspect that the degree to which various applications of anthropic reasoning are sensitive to that subjective element is inversely related to how scientifically robust those applications are. The most rigorous uses of anthropic reasoning have the property that they give the same result for almost any choice of reference class (satisfying only some very weak constraints).

In passing, we may note one interesting constraint on the choice of reference class. It turns out (for reasons that we do not have the space to elaborate on here) that a reference class definition according to which only *subjectively indistinguishable* observer-moments are placed in the same reference class is too narrow. (Two observer-moments are subjectively indistinguishable if they don't have any information that enables them to tell which one is which.) In other words, there are cases in which you should reason as if your current observer-moment were randomly selected from a class of observer-moments that includes ones of which you know that they are not your own current observer-moment. This fact makes anthropic reasoning a less simple affair than would otherwise have been the case.

The use of SSSA and the relativization of the reference class that SSSA enables thus seem to make it possible to coherently reject both the presumptuous philosopher's and the serpent's arguments, while at the same time one can show how to get plausible results in *Incubator* and several other

thought experiments as well as in various scientific applications, some of them novel. The theory can be condensed into one general formula: the Observation Equation, which specifies the probabilistic bearing on hypotheses of evidence that contains an indexical component.<sup>5</sup> Along with various constraints on permissible choices of reference classes, this forms the core of a theory of observation selection effects.

**§10.** AS A FINAL EXAMPLE, LET US CONSIDER AN EASY APPLICATION OF OBSERVATION selection theory to a puzzle that many drivers on the motorway may have wondered about (and cursed). Why is it that the cars in the other lane seem to be moving faster than you?

One might be inclined to account for the phenomenon by invoking Murphy's Law ("If anything can go wrong, it will," discovered by Edward A. Murphy, Jr, in 1949). However, a paper in *Nature* by Redelmeier and Tibshirani, published a couple of years ago, seeks a deeper explanation.<sup>6</sup> They present some evidence that drivers on Canadian roadways (where faster cars are not expected to move into more central lanes) think that the next lane is typically faster. They seek to explain the drivers' perceptions by appealing to a variety of psychological factors. For example:

"A driver is more likely to glance at the next lane for comparison when he is relatively idle while moving slowly;"

"Differential surveillance can occur because drivers look forwards rather than backwards, so vehicles that are overtaken become invisible very quickly, whereas vehicles that overtake the index driver remain conspicuous for much longer;"

"Human psychology may make being overtaken (losing) seem more salient than the corresponding gains."

The authors recommend that drivers should be educated about these effects in order to reduce the temptation to switch lanes repeatedly. This would reduce the risk of accidents, which are often caused by poor lane changes.

While all these psychological illusions might indeed occur, there is a more straightforward explanation for the drivers' persistent suspicion that cars in the next lane are moving faster. Namely, cars in the next lane actually do go faster!

One frequent cause of a lane (or a segment of a lane) being slow is that there are too many cars in it. Even if the ultimate cause is something else (for example, road work) there is nonetheless typically a negative correlation between the speed of a lane and how densely packed the vehicles driving in it are. This implies that a disproportionate fraction of the average driver's time is spent in slow lanes. If you think of your present observation, when you are driving on the motorway, as a random sample from all observations made by drivers, then chances are that your observation will be made from the viewpoint that most such observer-moments have, which is the view-

point of the slow-moving lane. In other words, appearances are faithful: more often than not, for most observer-moments, the “next” lane *is* faster.

Even when two lanes have the same average speed, it can be advantageous to switch lanes. For what is relevant to a driver who wants to reach her destination as quickly as possible is not the average speed of the lane as a whole, but rather the speed of some segment extending maybe a couple of miles forward from the driver’s current position. More often than not, the next lane has a higher average speed at this scale than does the driver’s present lane. On average, there is therefore a benefit to switching lanes (which of course has to be balanced against the costs of increased levels of effort and risk).

Adopting a thermodynamics perspective, it is also easy to see that (at least in the ideal case) increasing the “diffusion rate” (that is, the probability of lane-switching) will speed the approach to “equilibrium” (where there are equal velocities in both lanes), thereby increasing the road’s throughput and the number of vehicles that reach their destinations per unit time.

To summarize, in understanding this problem we must not ignore its inherent observation selection effect. This resides in the fact that if we randomly select an observer-moment of a driver and ask her whether she thinks the next lane is faster, more often than not we have selected an observer-moment of a driver who is in a lane which is in fact slower. When we realize this, we see that no case has been made for recommending that drivers change lanes less frequently.<sup>7</sup>

**§11.** OBSERVATION SELECTION THEORY (ALSO KNOWN AS ANTHROPIC REASONING), which aims to help us detect, diagnose, and cure the biases of observation selection effects, is a philosophical goldmine. Few branches of philosophy are so rich in empirical implications, touch on so many important scientific questions, pose such intricate paradoxes, and contain such generous quantities of conceptual and methodological confusion that need to be sorted out. Working in this area is a lot of intellectual fun.

The mathematics used in this field, such as conditional probabilities and Bayes’s theorem, are covered by elementary arithmetic and probability theory. The topic of observation selection effects *is* extremely complex, yet the difficulty lies not in the math, but in grasping and analyzing the underlying principles.  $\phi$

## *Notes*

<sup>1</sup>For further explorations of this and related principles, see Bostrom (1997), (2001), and (2002b).

<sup>2</sup>Lewis (1986)

<sup>3</sup>See, for example, Hall (1994), Lewis (1994), and Thau (1994).

<sup>4</sup>See Bostrom (2002a). Principles or forms of inferences that are similar to *SIA* have also been discussed by Dieks (1992), Smith (1994), Leslie (1996), Oliver and Korb (1997), Bartha and Hitchcock (1999) and (2000), and Olum (2002).

$${}^5 P_\alpha(h|e) = (1/\gamma) \sum_{\sigma \in \Omega_h \cap \Omega_e} \frac{P_\alpha(w_\sigma)}{|\Omega_\sigma \cap \Omega(w_\sigma)|} \quad (\text{Observation Equation})$$

Here,  $\alpha$  is the observer-moment whose subjective probability function is  $P_\alpha$ .  $\Omega_h$  is the class of all possible observer-moments about whom  $h$  is true;  $\Omega_e$  is the class of all possible observer-moments about whom  $e$  is true;  $\Omega_\alpha$  is the class of all observer-moments that  $\alpha$  places in the same reference class as herself;  $w_\alpha$  is the possible world in which  $\alpha$  is located; and  $\gamma$  is a normalization constant

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

The Observation Equation can be generalized to allow for different observer-moments within the reference class having different “weights,” an option that might be of relevance for instance in the context of the many-worlds version of quantum theory.

<sup>6</sup> Redelmeier and Tibshirani (1999).

<sup>7</sup> The above reasoning applies to a driver who is currently on the road wondering why she is in the slow lane. When considering the problem retrospectively, that is, when you are sitting at home thinking back on your experiences on the road, the situation is more complicated and requires also taking into account differential recall (a psychological factor that may make you more likely to remember and bring to mind certain kinds of experiences) and the fact that while the slow lane contains more *observer-moments*, it may nevertheless be true that more *drivers* have passed through the fast lane.

## Bibliography

- Bartha, P. and C. Hitchcock, “No One Knows the Date or the Hour: An Unorthodox Application of Rev. Bayes’s Theorem,” in *Philosophy of Science (Proceedings)* 66 (1999): S329-S353.
- Bartha, P. and C. Hitchcock, “The Shooting-Room Paradox and Conditionalizing on Measurably Challenged Sets,” in *Synthese* 108(3) (2000): 403-437.
- Bostrom, N., “Investigations into the Doomsday argument.” *Preprint* (1997). <<http://www.anthropic-principles.com/preprints/inv/investigations.html>>
- Bostrom, N., “The Doomsday argument, Adam & Eve, UN++, and Quantum Joe.” in *Synthese* 127(3) (2001): 359-387.
- Bostrom, N., *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (New York: Routledge, 2002a).
- Bostrom, N., “Self-Locating Belief in Big Worlds: Cosmology’s Missing Link to Observation,” in *Journal of Philosophy* 99(12) (2002 b).
- Dieks, D., “Doomsday—Or: the Dangers of Statistics,” in *Philosophical Quarterly* 42(166) (1992): 78-84.
- Hall, N., “Correcting the Guide to Objective Chance,” *Mind* 103(412) (1994): 505-517.
- Leslie, J., *The End of the World: The Science and Ethics of Human Extinction* (London: Routledge, 1996).
- Lewis, D., *Philosophical Papers* (New York: Oxford University Press, 1986).
- Lewis, D., “Humean Supervenience Debugged,” in *Mind* 103(412) (1994): 473-490.
- Oliver, J. and K. Korb, *A Bayesian analysis of the Doomsday Argument*, Department of Computer Science, Monash University, 1997.
- Olum, K., “The Doomsday Argument and the Number of Possible Observers,” in *Philosophical Quarterly* 52(207) (2002): 164-184.
- Redelmeier, D. A. and R. J. Tibshirani, “Why cars in the other lane seem to go faster,” in *Nature* 401 (1999): 35.
- Smith, Q., “Anthropic Explanations in Cosmology,” in *Australasian Journal of Philosophy* 72(3) (1994): 371-382.

Thau, M., “Undermining and Admissibility,” in *Mind* 103(412) (1994): 491-503.