The Doomsday Argument, Adam & Eve, UN^{++}, and Quantum Joe

Dr. Nick Bostrom

Department of Philosophy

Yale University

P. O. Box 208306, Yale Station

New Haven, Connecticut 06520

U. S. A.

Web: http://www.nickbostrom.com

Email: nick@nickbostrom.com

Tel: (203) 432-4771

Fax: (203) 432-1673

Running title: The Doomsday Argument *et al*.

The Doomsday Argument, Adam & Eve, UN^{++}, and Quantum Joe

ABSTRACT. The Doomsday argument purports to show that the risk of the human species going extinct soon has been systematically underestimated. This argument has something in common with controversial forms of reasoning in other areas, including: game theoretic problems with imperfect recall, the methodology of cosmology, the epistemology of indexical belief, and the debate over so-called fine-tuning arguments for the design hypothesis. The common denominator is a certain premiss: the Self-Sampling Assumption. We present two strands of argument in favor of this assumption. Through a series of thought experiments we then investigate some bizarre *prima facie* consequences – backward causation, psychic powers, and an apparent conflict with the Principal Principle.

The Self-Sampling Assumption and its use in the Doomsday argument

Let a person's *birth rank* be her position in the sequence of all observers who will ever have existed. For the sake of argument, let us grant that the human species is the only intelligent life form in the cosmos.[1] Your birth rank is then approximately 60 billionth, for that is the number of humans who have lived before you. The Doomsday argument proceeds as follows.

Compare two hypotheses about how many humans there will have been in total:[2]

h_1: = "There will have been a total of 200 billion humans."
h_2: = "There will have been a total of 200 trillion humans."

Suppose that after considering the various empirical threats that could cause human extinction (species-destroying meteor impact, nuclear Armageddon, self-replicating

nanobots destroying the biosphere, etc.) you still feel fairly optimistic about our prospects:

$\Pr(h_1) = .05$
$\Pr(h_2) = .95$

But now consider the fact that your birth rank is 60 billionth. According to the doomsayer, it is more probable that you should have that birth rank if the total number of humans that will ever have lived is 200 billion than if it is 200 trillion; in fact, your having that birth rank is one thousand times more probable given h_1 than given h_2:

$\Pr(\text{“My rank is 60 billionth.”} \mid h_1) = 1 / 200 \text{ billions}$
$\Pr(\text{“My rank is 60 billionth.”} \mid h_2) = 1 / 200 \text{ trillions}$

With these assumptions, we can use Bayes’s theorem to derive the posterior probabilities of h_1 and h_2 after taking your low birth rank into account:

plausibility than is the Doomsday argument. These other forms of reasoning include methods of deriving observational consequences from cosmological models (Leslie 1989; Bostrom 2000b; Bostrom 2001a), arguments concerning how improbable was the evolution of intelligent life on Earth (Carter 1983; Carter 1989), and game theoretic problems involving imperfect recall (e.g. Grove 1997; Piccione and Rubinstein 1997; Elga 2000). Tracking down the implications of this randomness principle has relevance for each of these domains.

The randomness assumption is invoked in the Doomsday argument in the step where the conditional probability of your having a specific birth rank given hypothesis *h* is set equal to the inverse of the number of observers that would exist if *h* were true. We can dub it the *Self-Sampling Assumption*:

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

For the purposes of this paper we can take the reference class to consist of all (intelligent) observers, although one of the lessons one might want to draw from our investigation is that this is too generous a definition and that the only way that SSA can continue to be defensibly used is by incorporating some restriction on the reference class such that not all observers are included in every observer's reference class.

However, even with the stipulation that we take the reference class to be the class of all observers, our formulation of SSA is still vague in that it leaves open at least two important questions: What counts as an observer? And what is the sampling density with which you have been sampled? How these areas of vagueness are resolved has serious consequences for what empirical predictions one gets when applying SSA to real empirical situations. Yet for present purposes we can sidestep these issues by introducing some simplifying assumptions. These will not change the fundamental principles involved but will on the contrary make it easier for us to focus on them.

To this effect, let's consider an imaginary world where there are no borderline cases of what counts as an observer and where the observers are sufficiently similar to each other to justify using a uniform sampling density (rather then one, say, where long-lived observers get a

proportionately greater weight). Thus, let us suppose for the sake of illustration that the only observers in existence are human beings, that we have no evolutionary ancestors, that all humans are fully self-aware and are knowledgeable about probability theory and anthropic reasoning, etc., that we all have identical life spans and that we are equal in any other arguably relevant respect. Assume furthermore that each human has a unique birth rank, and that the total number of humans that will ever have lived is finite.[4]

Under these assumptions, we get as a corollary of the SSA that

$$\text{(D)} \qquad \Pr(R = r \mid N = n) = \begin{cases} 1/n & \text{for } 0 \le r \le n \\ 0 & \text{for } r > n \end{cases},$$

where R and N are random variables: N representing the total number of people that will have lived, and R the birth rank of the particular person doing the reasoning. I call this expression "D" because of its complicity in the Doomsday argument. It is responsible for supplying the premiss from which the conditional probabilities (of you having a particular birth rank given a hypothesis about the duration of the human species) are derived. Without this premiss, the Doomsday argument could not get off the ground.

Arguments for the Self-Sampling Assumption

Before taking up the pursuit of some of the counterintuitive consequences of SSA, it is worth pausing to briefly consider some arguments that support SSA. These fall into two categories. First, there are a variety of thought experiments that describe situations in which it is plausible that one should reason in accordance with SSA. Second, there are arguments pointing to a methodological need for a principle like SSA in concrete scientific applications. It can be claimed, for example, that SSA serves to bridge a troublesome cleft between cosmological theory and observation.

The thought experiments that seem to favor adopting SSA include *The 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 the cells. What color should you think it is?

It seems that – in accordance with SSA – you should think that you are in a blue cell, with 90% probability. This answer is both intuitively plausible to many people and can be backed up by additional considerations. For instance, if all prisoners bet in accordance with SSA, then ninety per cent of then will win their bets; if you take part in great number of similar experiments, then you will likely in the long run find yourself in blue cells in ninety per cent of the cases; and so on. And the result doesn't seem to depend any assumptions about how the prisoners came to inhabit the cells they are in. Whether they were assigned to their cells by a random mechanism like lot or they were destined by physical laws to end up where they are, makes no difference so long as the prisoners aren't capable of figuring out their location from any knowledge they have of those circumstances. It is their subjective uncertainty that is guiding their credence assignments in *Dungeon*.

One might therefore think that the prisoners should assign credence in accordance with SSA only as long as they are uncertain about which cell they are in. Clearly, once you've stepped out of your cell and discovered that the outside is indeed blue, you should no longer assign a 90% credence to that hypothesis. Instead your credence is now unity (or very close to unity). Does this mean that you should reason in accordance with SSA only until such a time that you have direct empirical evidence as to what position you are in? If so, the SSA would not in any way support the Doomsday argument, since we have a lot of evidence that enables us to determine what our (approximate) birth ranks are in the human species. If you should cease to regard yourself a random sample once you have identifying information about the sample (yourself), then SSA would amount to nothing but a restricted and fairly toothless version of the principle of indifference, and it would not have any of the counterintuitive consequences that we will encounter later in this paper.

That reading of SSA is not what the doomsayers have in mind, however. To see what's at stake, rather, consider the following thought experiment:

The Incubator

Stage (a): The world consists of a dungeon with one hundred cells. The outside of each cell has a unique number painted on it (which can't be seen from the inside); the numbers being the integers between 1 and 100. The world also contains a mechanism which we can term the *incubator*.[5] The incubator first creates one observer in cell #1. It then activates a randomization mechanism; let's say it flips a fair coin. If the coin falls tails, the incubator does nothing more. If the coin falls heads, the incubator creates one observer in each of the cells ##2 - 100. Apart from this, the world is empty. It is now a time well after the coin has been tossed and any resulting observers have been created. Everyone knows all the above.

Stage (b): A little later, you have just stepped out of your cell and discovered that it is #1.

Here the suggestion is that at stage (a) you should assign a 50% probability to the coin having landed heads. Moreover, your conditional probabilities at stage (a) of being in a particular cell, say cell #1, given that the coin fell heads seems to be 1%, since if the coin fell heads then there are one hundred people, any one of which might be you for all you know, and only one of which is in cell #1. Similarly, your conditional probability of being in cell #1 given that the coin fell tails is 100%, since that's the only place you could be given that outcome. This is in accordance with SSA.

What you should think at stage (b) seems to follow from this. If you continue to accept a prior probability of heads equal to 50%, and conditional probabilities of being in cell #1 equal to 1% (or 100%) given Heads (or Tails), then it follows from Bayes's theorem that after finding that you are in cell #1 in order to be coherent you must assign a posterior probability of Heads that is equal to 1/101, and a posterior probability of Tails that is equal to 100/101.[6] In other words, you go from being completely ignorant about how the coin landed (50% probability of Tails) to being quite confident that it landed tails (99% probability).[7]

In this reasoning you continue to regard yourself as a random sample throughout. This is analogous to a case where you draw a random sample from an urn which contains either one ball that is numbered "#1", or one hundred balls which are numbered consecutively from "#1" to

"#100". Suppose a fair coin toss determined which of these alternatives obtains, so the prior probability of the urn containing only one ball is 50%. (Let's say Tails gives one ball, Heads a hundred.) The probability that the ball you have drawn is #1 is $P = (½ \times 1/100) + (½ \times 1) = 50.5\%$. After you've examined the ball and found that it is #1, it remains correct to view the ball as a random sample, and of course that doesn't mean that you should continue to assign a 50.5% probability to it being #1. Rather, you simply add in the new information you've obtained about the random sample and update your beliefs accordingly. It remains the case, for example, that the conditional probability of the ball (the random sample) being #1 is much greater given Tails than Heads, and you can use this to infer, after finding that you drew ball #1, that the coin probably fell tails and the urn contained only one ball. In the same manner, the doomsayer maintains that we should regard ourselves as random samples even though we know many facts that show that we are a product of our age and that tie us to a specific position in the human species.

The injunction that we should reason as if we were random observers is a methodological prescription about what values to give to certain conditional probabilities, in particular those of the form:

Pr("*I'm* an observer with such and such properties." | "The world is such and such.")[8]

This methodological prescription is intended as an epistemological principle that is independent of any assumptions about us having been generated through some objectively random process. There is no need to assume a time-traveling stork that had an equal probability of dropping you off at any location throughout history where a human child was about to be delivered.

Now to the second kind of arguments for SSA. These are arguments that point to legitimate scientific needs that rely on the services provided by SSA. We can most readily see this in cosmology, where the basic idea is as follows. It seems that the cosmos is very big, so big in fact that we have reason to believe that every possible observation is made.[9] How can we ever test theories which say that the cosmos is that big? For any observation we specify, such theories assign a very high probability (a probability of one in the case of typical infinite-cosmos theories) to the hypothesis that that observation is made. So all such theories seem to be perfectly probabilistically compatible with every possible observation; from which it would follow that

empirical evidence cannot possibly give us any reason whatever for favoring one such infinite-cosmos theory to another.[10] Even a theory saying that, say, the gravitational constant has a different value than the one we have observed would not be in any way disfavored by our observations, because even on the theory with the deviant value of the gravitational constant, observations like ours would be made, with probability one.

This line of reasoning must be faulty, for cosmologists are constantly testing and revising big-cosmos theories in light of new empirical evidence. The way to resolve the conundrum, it seems, is by insisting that when evaluating a theory in light of empirical evidence, we should use the most specific version of the evidence that is known. And in this case, we know more than that some observation *b* has been made. We know that *b* has been made *by us*. The question thus arises, how probable was it on rival theories that *we* should make that particular observation? This is where SSA comes in. According to SSA, we should reason as if we were random observers. Using this principle, we can then infer that the conditional probability (given theory *T*) of a specific observer making observation *b* should be set equal to the expected *fraction* of all observers who (according to *T*) make observation *b*. SSA enables us to take this step from fractions to probabilities. By doing so, SSA rescues us from a methodological embarrassment, and it deserves credit for that.[11]

SSA is thus not an arbitrary or silly assumption pulled from an empty hat. It is a methodological principle supported by two fairly compelling strands of argument. This, in addition to its role in the Doomsday argument, makes it important to learn that SSA comes with a price tag attached: adopting it commits one to certain consequences which one might feel are unacceptable. Being clear about this will help us be more informed if we decide to search for a more affordable substitute for SSA.

The Adam & Eve experiments

The three *Adam & Eve* thought experiments that follow are all variations on the same theme. They put different problematic aspects of SSA into focus.

First experiment: Serpent's Advice

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.[12] 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's 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!"

Given SSA and the stated assumptions, it is easy to see that the serpent's argument is sound. We have $\Pr(R \leq 2 | N = 2) = 1$ and using SSA, $\Pr(R \leq 2 | N > 2 \cdot 10^9) < 10^{-9}$. We can assume that the prior probability of getting pregnant (based on ordinary empirical considerations) after congress is very roughly one half, $\Pr(N = 2) \approx \Pr(N > 2 \cdot 10^9) \approx .5$. Thus, according to Bayes's theorem we have

$$\begin{aligned} &\Pr(N > 2 \cdot 10^9 | R \leq 2) \\ &= \frac{\Pr(R \leq 2 | N > 2 \cdot 10^9) \Pr(N > 2 \cdot 10^9)}{\Pr(R \leq 2 | N > 2 \cdot 10^9) \Pr(N > 2 \cdot 10^9) + \Pr(R \leq 2 | N = 2) \Pr(N = 2)} \\ &< 10^{-9} \end{aligned}$$

Eve has to conclude that the risk of her getting pregnant is negligible.

This result is counterintuitive. Most people's intuition, at least at first glance, is that it would be irrational for Eve to think that the risk is that low. It seems foolish of her to act as if she were extremely unlikely to get pregnant – it seems to conflict with empirical data. And we can assume she is fully aware of these data, at least to the extent to which they are about past events. We can assume that she has access to a huge pool of statistics, maybe based on some population

of lobotomized human drones (lobotomized so that they don't belong to the reference class, the class from which Eve should consider herself a random sample). Yet all this knowledge, combined with everything there is to know about the human reproductive system, would not change the fact that it would be irrational for Eve to believe that the risk of her getting pregnant is anything other than effectively nil. This is a strange result, but it follows from SSA.[13]

Second experiment: Lazy Adam

The next example effects another turn of the screw, deriving a consequence that has an even greater degree of initial counterintuitiveness:

> Assume as before that Adam and Eve were once the only people and that they know for certain that if they have a child they will be driven out of Eden and will have billions of descendants. But this time they have a foolproof way of generating a child, perhaps using advanced *in vitro* fertilization. Adam is tired of getting up every morning to go hunting. Together with Eve, he devises the following scheme: *They form the firm intention that unless a wounded deer limps by their cave, they will have a child*. Adam can then put his feet up and rationally expect with near certainty that a wounded dear – an easy target for his spear – will soon stroll by.

One can verify this result the same way as above, choosing appropriate values for the prior probabilities. The prior probability of a wounded deer limping by their cave that morning is one in ten thousand, say.

In the first experiment we had an example of what looked like anomalous precognition. Here we also have (more clearly than in the previous case) the appearance of psychokinesis. If the example works, which it does if we assume SSA, it almost seems as if Adam is *causing* a wounded deer to walk by. For how else could one explain the coincidence? Adam knows that he can repeat the procedure morning after morning and that he should expect a deer to appear each time. Some mornings he may not form the relevant intention and on those mornings no deer turns up. It seems too good to be mere chance; Adam is tempted to think he has magical powers.

Third experiment: Eve's Card Trick

> One morning, Adam shuffles a deck of cards. Later that morning, Eve, having had no contact with the cards, decides to use her willpower to retroactively choose what card lies top. She decides that it shall have been the dame of spades. In order to ordain this outcome, Eve and Adam form the firm intention to have a child unless the dame of spades is top. They can then be virtually certain that when they look at the first card they will indeed find the dame of spades.

Here it looks as if the couple is in one and the same act performing both psychokinesis and backward causation. No mean feat before breakfast.

These three thought experiments seem to show that SSA has bizarre consequences: strange coincidences, precognition, psychokinesis and backward causation in situations where we would not expect such phenomena. If these consequences are genuine, they must surely count heavily against the unrestricted version of SSA, with ramifications for the Doomsday argument and other forms of anthropic reasoning that rely on that principle.

However, we shall now see that such an interpretation misreads the experiments. The truth is more interesting than that. A careful look at the situation reveals that SSA, in subtle ways, wiggles its way out of the worst of the imputed implications.

Analysis of *Lazy Adam*: predictions and counterfactuals

This section discusses the second experiment, *Lazy Adam*. I think that the first and the third experiments can be analyzed along similar lines.

Adam can repeat the *Lazy Adam* experiment many mornings.[14] And the experiment seems *prima facie* to show that, given SSA, there will be a series of remarkable coincidences between Adam's procreational intentions and appearances of wounded deer. It was suggested that such a series of coincidences could be a ground for attributing paranormal causal powers to Adam.

The inference from a long series of coincidences to an underlying causal link can be disputed. Whether such an inference is legitimate would depend on how long is the series of

coincidences, what are the circumstances, and also on what theory of causation one should hold. If the series were sufficiently long and the coincidences sufficiently remarkable, intuitive pressure would mount to give the phenomenon a causal interpretation; and one can fix the thought experiment so that these conditions are satisfied. For the sake of argument, we may assume the worst case for SSA, namely that if the series of coincidences occurs then Adam has anomalous causal powers. I shall argue that even if we accept SSA, we can still think that neither strange coincidences nor anomalous causal powers would have existed if the experiment had been carried out.

We need to be careful when stating what is implied by the argument given in the thought experiment. All that was shown is that Adam would have reason to believe that his forming the intentions will have the desired outcome. The argument can be extended to show that Adam would have reason to believe that the procedure can be repeated: provided he keeps forming the right intentions, he should think that morning after morning, a wounded deer will turn up. If he doesn't form the intention on some mornings, then on those mornings he should expect deer *not* to turn up. Adam thus has reason to think that deer turn up on those and only on those mornings for which he formed the relevant intention. In other words, Adam has reason to believe there will be a coincidence. However, we cannot jump from this to the conclusion that there will actually be a coincidence. Adam could be mistaken. And he could be mistaken even though he is (as the argument in *Lazy Adam* showed, assuming SSA) perfectly rational.

Imagine for a moment that you are looking at the situation from an external point of view. That is, suppose (*per impossible?*) that you are an intelligent observer who is not a member of the reference class. Suppose you know the same non-indexical facts as Adam; that is, you know the same things as he does except such things as that "I am Adam" or "I am among the first two humans" etc. Then the probability you should assign to the proposition that a deer will limp by Adam's cave one specific morning conditional on Adam having formed the relevant intention earlier that morning is the same as what we called Adam's prior probability of deer walking by – one in ten thousand. As an external observer you would consequently not have reason to believe that there were to be a coincidence.[15]

Adam and the external observer, both being rational but having different information, make different predictions. At least one of them must be mistaken (although both are "right" in

the sense of doing the best they can with the evidence available to them). In order to determine who was in fact mistaken, we should have to decide whether there would be a coincidence or not. Nothing said so far settles this question. There are possible worlds where a deer does turn up on precisely those mornings when Adam forms the intention, and there are other possible worlds with no such coincidence. The description of the thought experiment does not specify which of these two kinds of possible worlds we are referring to; it is underdetermined in this respect.

So far so good, but we want to be able to say something stronger. Let's pretend there actually once existed these two first people, Eve and Adam, and that they had the reproductive capacities described in the experiment. We would want to say that if the experiment had actually been done (i.e. if Adam had formed the relevant intentions on certain mornings) then almost certainly *he would have found no coincidence*. Almost certainly, no wounded deer would have turned up. That much seems common sense. If SSA forced us to relinquish that conviction, it would count quite strongly as a reason for rejecting SSA.

We therefore have to evaluate the counterfactual: *If Adam had formed the relevant intentions, would there have been a coincidence?* To answer this, we need a theory of conditionals. I will use a simplified version of David Lewis' theory[16] but I think what I will say generalizes to other accounts of conditionals. Let *w* denote the actual world. (We are pretending that Adam and Eve actually existed and that they had the appropriate reproductive abilities etc.) To determine what would have happened had Adam formed the relevant intentions, we look at the closest[17] possible world *w'* where he did do the experiment. Let *t* be the time when Adam would have formed the intentions. When comparing worlds for closeness to *w*, we are to disregard features of them that exclusively concern what happens after *t*. Thus we seek a world in which Adam forms the intentions and which is maximally similar to *w* in two respects: first, in its history up to *t*; and, second, in its laws. Is the closest world (*w'*) to *w* on these accounts and where Adam forms the intentions a world where deer turn up accordingly, or is it a world where there is no Adam-deer correlation?

The answer is quite clearly that there is no Adam-deer correlation in *w'*. For such a *w'* can be more similar to *w* on both accounts than can any world containing the correlation. Regarding the first account, whether there is a coincidence or not in a world presumably makes little difference as to how similar it can be to *w* with respect to its history up to *t*. But what

difference it makes is in favor of no coincidence. This is so because in the absence of a correlation the positions and states of the deer in the neighborhood, at or shortly before *t*, could be exactly as in *w* (where none happened to stroll past Adam's cave on the mornings when he did the experiment). The presence of a correlation, on the other hand, would entail a world that would be somewhat different regarding the initial states of the deer.

Perhaps more decisively, a world with no Adam-deer correlation would tend to win out on the second account as well. *w* doesn't (as far as we know) contain any instances of anomalous causation. The laws of *w* do not support anomalous causation. The laws of any world containing an Adam-deer correlation, at least if the correlation were of the sort that would prompt us to ascribe it to an underlying causal connection, include laws supporting anomalous causation. By contrast, the laws of a world lacking the Adam-deer correlation could easily have laws exactly as in *w*. Similarity of laws would therefore also favor a *w'* with no correlation.

Since there is no correlation in *w'*, the following statement is true: "If Adam had formed the intentions, he would have found no correlation". Although Adam would have reason to think that there would be a coincidence, he would find he was mistaken.

One might wonder: if *we* know all this, why can't Adam reason in the same way? Couldn't he too figure out that there will be no coincidence?

He couldn't, and the reason is that he is lacking some knowledge you and I have. Adam has no knowledge of the future that will show that his creative hunting technique will fail. If he does his experiment and deer do turn up on precisely those mornings he forms the intention, then it could (especially if the experiment were successfully repeated many times) be the case that the effect should be ascribed to a genuine psychokinetic capacity. If he does the experiment and no deer turns up, then of course he has no such capacity. But he has no means of knowing that no deer turns up. The evidence available to him strongly favors the hypothesis that there *will* be a coincidence. So although Adam may understand the line of reasoning that we have been pursuing here, it will not lead him to the conclusion we arrived at, because he lacks a crucial premiss.

There is a puzzling point here that needs be addressed. Adam knows that if he forms the intentions then he will very likely witness a coincidence. But he also knows that if he doesn't form the intentions then it will be the case that he will live in a world like *w*, where it is true that

had he done the experiment he would most likely *not* have witnessed a coincidence. That looks paradoxical. Adam's forming (or not forming) the conditional procreational intentions gives him relevant information. Yet, the only information he gets is about what choice he made. If that information makes a difference as to whether he should expect to see a coincidence, isn't that just to say that his choice affects whether there will be a coincidence or not? If so, it would seem he has got paranormal powers after all.

A more careful analysis reveals that this conclusion doesn't follow. True, the information Adam gets when he forms the intentions is about what choice he made. This information has a bearing on whether to expect a coincidence or not, but that doesn't mean that the choice is a *cause* of the coincidence. It is simply an *indication* of a coincidence. Some things are good indicators of other things without causing them. Take the stock example: the barometer's falling may be a good indicator of impending rain, if you knew something about how barometers work, but it is certainly not a cause of the rain. Similarly, there is no need to think of Adam's decision to procreate if and only if no deer walks by as a *cause* of that event, although it will lead Adam to rationally believe that that event will happen.

One may feel that an air of mystery lingers on. Maybe we can put it into words as follows: Let E be the proposition that Adam forms the reproductive intention at time $t = 1$, let C stand for the proposition that there is a coincidence at time $t = 2$ (i.e. that a deer turns up). It would seem that the above discussion commits one to the view that at $t = 0$ Adam knows (probabilistically) the following:

(1) If E then C.

(2) If $\neg E$ then $\neg C$.

(3) If $\neg E$ then "if E then it would have been the case that $\neg C$".

And there seems to be a conflict between (1) and (3).

I suggest that the appearance of a conflict is due to an equivocation in (3). To bring some light into this, we can paraphrase (1) and (2) as:

(1') $\Pr_{\text{Adam}}(C|E) \approx 1$

(2') $\Pr_{\text{Adam}}(\neg C|\neg E) \approx 1$

But we cannot paraphrase (3) as:

(3') $\Pr_{\text{Adam}}(\neg C|E) \approx 1$

When I said earlier, "If Adam had formed the intentions, he would have found no correlation", I was asserting this on the basis of background information that is available to us but not to Adam. Our set of background knowledge differs from Adam's in respect to both non-indexical facts (we have observed the absence of any subsequent correlation between peoples' intentions and the behavior of deer) and indexical facts (we know that we are not among the first two people). Therefore, if (3) is to have any support in the preceding discussion, it should be explicated as:

(3'') $\Pr_{\text{We}}(\neg C|E) \approx 1$

This is not in conflict with (1'). I also asserted that Adam could know this. This gives:

(4) $\Pr_{\text{Adam}}(\text{“}\Pr_{\text{We}}(\neg C|E) \approx 1\text{”}) \approx 1$

At first sight, it might seem as if there is a conflict between (4) and (1). However, appearances in this instance are deceptive.

Let's first see why it could *appear* as if there is a conflict. It has to do with the relationship between $\Pr_{\text{Adam}}$ and $\Pr_{\text{We}}$. We have assumed that $\Pr_{\text{Adam}}$ is a rational probability assignment (in the sense: not just coherent but "reasonable, plausible, intelligent" as well) relative to the set of background knowledge that Adam has at $t = 0$. And $\Pr_{\text{We}}$ is a rational probability assignment relative to the set of background knowledge that we have, say at $t = 3$. (And of course we pretend that we know that there actually was this fellow Adam at $t = 0$ and that he had the appropriate reproductive abilities etc.) But now, if we know everything Adam knew, and if in addition we have some extra knowledge, *and if Adam knows that*, then it is irrational of him to persist in believing what he believes. Instead he ought to adopt our beliefs, which he knows are

based on more information. At least this follows if we assume, as we may in this context, that our a priori probability function is identical to Adam's, and that we haven't made any computational error, and that Adam knows all this. That would then imply (3') after all, which contradicts (1').

The fallacy in this argument is that it assumes that Adam knows that we know everything he knows. Adam doesn't know that, because *he doesn't know that we exist*. He may well know that *if* we exist then we will know everything (at least every objective – non-indexical – piece of information) that he knows and then some. But as far as he is concerned, we are just hypothetical beings.[18] So all that Adam knows is that there is some probability function, the one we denoted Pr_{We}, that gives a high conditional probability of $\neg C$ given E. That gets him nowhere. There are infinitely many probability functions, and not knowing that we will actually exist he has no more reason to tune his own credence to our probability function than to any other.

To summarize the results so far, what we have shown is the following: Granting SSA, we should think that if Adam and Eve had carried out the experiment, there would almost certainly *not* have been any strange coincidences. There is thus no reason to ascribe anomalous causal powers to Adam. Eve and Adam would rationally think otherwise but they would simply be mistaken. Although they can recognize the line of reasoning we have been pursuing they won't be moved by its conclusion, because it hinges on a premiss that we – but not they – know is true. Good news for SSA.

One more point needs to be addressed in relation to *Lazy Adam*. We have seen that what the thought experiments demonstrate is not strange coincidences or anomalous causation but simply that Adam and Eve would be misled. Now, there might be a temptation to see this by itself as a ground for rejecting SSA – if a principle misleads people it is not reliable and should not be adopted. However, this temptation is to be resisted. There is a good answer available to the SSA-proponent, as follows: It is in the nature of probabilistic reasoning that some people using it, if they are in unusual circumstances, will be misled. Eve and Adam were in highly unusual circumstances – they were the first two humans – so we shouldn't bee too impressed by the fact that the reasoning based on SSA didn't work for them. For a fair assessment of the reliability of SSA we have to look at how it performs not only in exceptional cases but in more normal cases as well.

Compare the situation to the *Dungeon* gedanken. There, remember, one hundred people were placed in different cells and were asked to guess the color of the outside of their own cell. Ninety cells were blue and ten red. SSA recommended that a prisoner thinks that with 90% probability he is in a blue cell. If all prisoners bet accordingly, 90% of them will win their bets. The unfortunate 10% who happen to be in red cells lose their bets, but it would be unfair to blame SSA for that. They were simply unlucky. Overall, SSA leads 90% to win, compared to merely 50% if SSA is rejected and people bet at random. This consideration works in favor of SSA.

What about the "overall effect" of everybody adopting SSA in the three experiments pondered above? Here the situation is more complicated because Adam and Eve have much more information than the people in the cells. Another complication is that these are stories where there are two competing hypotheses about the total number of observers. In both these respects the thought experiments are similar to the Doomsday argument and presumably no easier to settle. What we are interested in here is finding out whether there are some *other* problematic consequences of SSA which are not salient in the Doomsday argument – such as strange coincidences and anomalous causation.

The UN^{++} gedanken: reasons, abilities, and decision theory

We shall now discuss a thought experiment which is similar to the Adam & Eve experiments but differs in that it is one that we might actually one day be able to carry out.

UN^{++}

It is the year 2100 A.D. and technological advances have enabled the formation of an all-powerful and extremely stable world government, UN^{++}. Any decision about human action taken by the UN^{++} will certainly be implemented. However, the world government does not have complete control over natural phenomena. In particular, there are signs that a series of *n* violent gamma ray bursts is about to take place at uncomfortably close quarters in the near future, threatening to damage (but not completely destroy) human settlements. For each hypothetical gamma ray burst in this series, astronomical observations give a 90% chance of it coming about. However, UN^{++} raises to the

occasion and passes the following resolution: It will create a list of hypothetical gamma ray bursts, and for each entry on this list it decides that if the burst happens, it will build more space colonies so as to increase the total number of humans that will ever have lived by a factor of *m*. By arguments analogous to those in the earlier thought experiments, UN^{++} can then be confident that the gamma ray bursts will not happen, provided *m* is sufficiently great compared to *n*.

The UN^{++} experiment introduces a new difficulty. For although creating UN^{++} and persuading it to adopt the plan would no doubt be a daunting undertaking, it is the sort of project that we could quite conceivably carry out by non-magical means. The UN^{++} experiment places *us* in more or less the same situation as Adam and Eve in the other three experiments. This twist compels us to carry the investigation one step further.

Let us suppose that if there is a long series of coincidences ("*C*") between items on the UN^{++} target list and failed gamma ray bursts then there is anomalous causation ("*AC*"). This supposition is more problematic than the corresponding assumption when we were discussing Adam and Eve. For the point of UN^{++} experiment is that it is claiming some degree of practical possibility, and it is not clear that this supposition could be satisfied in the real world. It depends on the details and o wprac+sm02 0 bly17 ect tearli2(o.5(be.3(d)3.4(tigatilsupposiFor the po-Tc(+sm0TJ-15.6885

and make it adopt the plan; and we have good reason (given SSA) to think that if we do this then there will be *C* and hence *AC*. But if we now have the *ability* to bring about *AC* then *we now, ipso facto, have AC*. Since this is absurd, we should reject SSA.

This reasoning is fallacious. Our forming UN^{++} and making it adopt the plan would be an *indication* to us that there is a correlation between the list and gamma ray bursts.[19] But it would not *cause* there to be a correlation unless we do in fact have *AC*. If we don't have *AC* then forming UN^{++} and making it adopt the plan (call this event "*A*") has no influence whatever on astronomical phenomena, although it misleads us to thinking we have. If we do have *AC* of the relevant sort, then of course the same actions would influence astronomical phenomena and cause a correlation. But the point is this: the fact that we have the ability to do *A* does not in any way determine whether we have *AC*. It doesn't even imply that we have reason to think that we have *AC*.

In order to be perfectly clear about this point, let me explicitly write down the inference I am rejecting. I'm claiming that from the following two premises:

(5) We have strong reasons to think that if we do *A* then we will have brought about *C*.

(6) We have strong reasons to think that we have the power to do *A*.

one cannot legitimately infer:

(7) We have strong reasons to think that we have the power to bring about *C*.

My reason for rejecting this inference is that one can consistently hold the conjunction of (5) and (6) together with the following:

(8) If we don't do *A* then the counterfactual "Had we done *A* then *C* would have occurred" is false.

There might be a temptation to think that the counterfactual in (8) would have been true even if don't do *A*. I suggest that this is due to the fact that (granting SSA) our conditional

probability of C given that we do A is large. Let's abbreviate this conditional probability 'Pr($C|A$)'. If Pr($C|A$) is large, doesn't that mean that C would (probably) have happened if we had done A? Not so. One must not confuse the conditional probability Pr($C|A$) with the counterfactual "C would have happened if A had happened". For one thing, the reason why your conditional probability Pr($C|A$) is large is that you have included indexical information (about your birth rank) in the background information. Yet one may well choose to exclude indexical information from the set of facts upon which counterfactuals are to supervene. (Especially so if one intends to use counterfactuals to define causality, which should presumably be an objective notion and therefore independent of indexical facts – see the next section for some further thoughts on this.)

So, to reiterate, even though Pr($C|A$) is large (as stated in (5)) and even though we can do A (as stated in (6)), we still know that, *given that we don't do A*, C almost certainly does not happen and would not have happened even if we had done A. As a matter of fact, we have excellent grounds for thinking that we won't do A. The UN^{++} experiment, therefore, does not show that we have reason to think that there is AC. Good news for SSA, again.

Finally, although it may not be directly relevant to assessing whether SSA is true, it is interesting to ask: *Would it be rational* (given SSA) *for* UN^{++} *to adopt the plan?*[20]

The UN^{++} should decrease its credence of the proposition that a gamma ray burst will occur if it decides to adopt the plan. Its conditional credence Pr(Gamma ray burst | A) is smaller than Pr(Gamma ray burst); that is what the thought experiment showed. Provided a gamma ray burst has a sufficiently great negative utility, non-causal decision theories would recommend that we adopt the plan if we can.

What about causal decision theories? If our theory of causation is one on which no AC would be involved even if C happens, then obviously causal decision theories would say that the plan is misguided and shouldn't be adopted. The case is more complicated on a theory of causation that says that there is AC if C happens. UN^{++} should then believe the following: If it adopts the plan, it will have caused the outcome of averting the gamma ray burst; if it doesn't adopt the plan, then it is not the case that had it adopted the plan it would have averted the gamma ray bursts. (This essentially just repeats (5) and (8).) The question is whether causal decision theories would under these circumstances recommend that UN^{++} adopt the plan.

The decision that UN^{++} makes gives it information about whether it has *AC* or not. Yet, when UN^{++} deliberates on the decision, it can only take into account information available to it prior to the decision, and this information doesn't suffice to determine whether it has *AC*. UN^{++} therefore has to make its decision under uncertainty. Since on a causal decision theory UN^{++} should do *A* only if it has *AC*, UN^{++} would have to act on some preliminary guess about how likely it seems that *AC*; and since *AC* is strongly correlated with what decision UN^{++} makes, it would also base its decision, implicitly at least, on a guess about what its decision will be. If it thinks it will eventually choose to do *A*, it has reason to think it has *AC*, and thus it should do *A*. If it thinks it will eventually choose not to do *A*, it has reason to think that it hasn't got *AC*, and thus should not do *A*. UN^{++} therefore is faced with a somewhat degenerate decision problem in which it should choose whatever it initially guesses it will come to choose. More could no doubt be said about the decision theoretical aspects of this scenario[21], but we will leave it at that.

Quantum Joe: SSA and the Principal Principle

Our final thought experiment probes the connection between SSA and objective chance:

> *Quantum Joe*
>
> Joe, the amateur scientist, has discovered that he is alone in the cosmos so far. He builds a quantum device which according to quantum physics has a one-in-ten chance of outputting any single-digit integer. He also builds a reproduction device which when activated will create ten thousand clones of Joe. He then hooks up the two so that the reproductive device will kick into action unless the quantum device outputs a zero; but if the output is a zero, then the reproductive machine will be destroyed. There are not enough materials left for Joe to reproduce in some other way, so he will then have been the only observer.

We can assume that quantum physics correctly describes the objective chances associated with the quantum device, and that Everett-type interpretations (including the many-worlds and the many-minds interpretations) are false; and that Joe knows this. Using the same kinds of argument

as before, we can show that Joe should expect that a zero come up, even though the objective (physical) chance is a mere 10%.

Our reflections on the *Adam & Eve* and UN^{++} apply to this gedanken also. But here we shall focus on another problem: the apparent conflict between SSA and David Lewis's Principal Principle.

The Principal Principle requires, roughly, that one proportion one's credence in a proposition B in accordance with one's estimate of the objective chance that B will come true (Lewis 1980; Mellor 1971). For example, if you know that the objective chance of B is x%, then your subjective credence of B should be x%, provided you don't have "inadmissible" information. An early formalization of this idea turned out to be inconsistent when applied to so-called "undermining" futures, but this problem has recently been solved through the introduction of the "new Principal Principle", which states that:

$$\Pr(B|HT) = \mathrm{Ch}(B|T)$$

H is a proposition giving a complete specification of the history of the world up to time t, T is the complete theory of chance for the world (giving all the probabilistic laws), Pr is a rational credence function, and Ch is the chance function specifying the world's objective probabilities at time t. (For an explanation of the *modus operandi* of this principle and of how it can constitute the centerpiece of an account of objective chance, see Lewis 1994; Thau 1994; Hall 1994.)

Now, Quantum Joe knows all the relevant aspects of the history of the world up to the time when he is about to activate the quantum device. He also has complete knowledge of quantum physics, the correct theory of chance for the world in which he is living. If we let B be the proposition that the quantum device outputs a zero, the new Principal Principle thus seems to recommend that he should set his credence of B equal to $\mathrm{Ch}(B|T) \approx 1/10$. Yet the SSA-based argument shows that his credence should be ≈ 1. Does SSA therefore require that we give up the Principal Principle?

I think this can be answered in the negative, as follows. True, Joe's credence of getting a zero should diverge from the objective chance of that outcome, even though he knows what that chance is. But that is because he is basing his estimation on inadmissible information. That being

so, the new Principal Principle does not apply to Joe's situation. The inadmissible information is indexical information about his Joe's own position in the human species. Normally, indexical information does not affect one's subjective credence in propositions whose objective chances are known. But in certain kinds of cases, such as the one we are dealing with here, indexical information turns out to be relevant and must be factored in.

It not really surprising that the Principal Principle, which expresses the connection between objective chance and rational subjective credence, is trumped by other considerations in cases like these. For objective chances can be seen as concise, informative summaries of patterns of local facts about the world. (That is certainly how they are seen in Lewis's analysis.) But the facts that form the supervenience base for chances are rightly taken not to include indexical facts, for chances are meant to be objective. Since indexical information is not baked into chances, it is only to be expected that your subjective credence may have to diverge from known objective chances if you have additional information of an indexical character that needs be taken into account.

So Quantum Joe can coherently believe that the objective chance (as given by quantum physics) of getting a zero is 10% and yet set his credence in that outcome close to one; he can accept both the Principal Principle and SSA.

Conclusion

SSA is a central premiss in the Doomsday argument. We have considered two strands of argument that support SSA: one based on thought experiments where many people have intuitions that lead to conclusions parallel to that of the Doomsday argument; the other based on the scientific need for a methodological principle that can establish a link between big-world cosmologies and observational consequences – a role that SSA is able to fill. These arguments establish at least that SSA deserves serious attention. It behooves anybody who would reject SSA to show why these arguments fail, and to propose a better principle in its stead.

We then turned to consider some challenges to SSA. In *Lazy Adam*, it looked as though on the basis of SSA we should think that Adam had the power to produce anomalous coincidences by will, exerting a psychokinetic influence on the nearby deer population. On closer

inspection, it turned out that SSA implies no such thing. It gives us no reason to think that there would have been coincidences or psychic causation if Adam had carried out the experiment. SSA does lead Adam to think otherwise, but he would simply have been mistaken. We argued that the fact that SSA would have misled Adam is no good argument against SSA. For it is in the nature of probabilistic reasoning that exceptional users will be misled, and Adam is such a user. To assess the reliability of SSA-based reasoning one has to look at not only the special cases where it fails but also the normal cases where it succeeds. We noted that in the *Dungeon* experiment, SSA maximizes the fraction of observers who are right.

With the UN^{++} gedanken, the scene was changed to one where we ourselves might actually have the ability to step into the role of Adam. We found that SSA does not give us reason to think that there will be strange coincidences or that we (or UN^{++}) have anomalous causal powers. However, there are some hypothetical (empirically implausible) circumstances under which SSA *would* entail that we had reason to believe these things. *If* we knew for certain that UN^{++} existed and had the power to create observers in the requisite numbers and possessed sufficient stability to certainly follow through on its original plan, and that the other presuppositions behind the thought experiment were also satisfied – no extraterrestrials, all observers created are in the reference class, etc. – *then* SSA implies that we should expect to see strange coincidences, namely that the gamma ray bursts on the UN^{++} target list would fizzle. (Intuitively: because this would make it enormously much less remarkable that we should have the birth ranks we have.) But we should think it extremely unlikely that this situation will arise.[22]

Finally, in *Quantum Joe* we examined an ostensible conflict between SSA and the Principal Principle. It was argued that this conflict is merely apparent because the SSA-line of reasoning relies on indexical information that should properly be regarded as "inadmissible" and thus outside the scope of the Principal Principle.

These triumphs notwithstanding, it is fair to characterize the SSA-based advice to Eve, that she need not worry about pregnancy, and its recommendation to Adam, that he should expect a deer to walk by given that the appropriate reproductive intentions are formed, and Quantum Joe's second-guessing of quantum physics, as deeply counterintuitive results. We are forced to espouse these implications if we accept the version of SSA discussed in this paper. Maybe the

lesson is that we should search for a version of SSA that avoids these consequences.[23] Thus modifying SSA may pull the rug from under the Doomsday argument.

Acknowledgements

I'm grateful for interesting discussions with Craig Callender, Milan M. Ćirković, Dennis Dieks, William Eckhardt, Adam Elga, Paul Franceschi, Mark Greenberg, Colin Howson, John Leslie, Peter Milne, Ken Olum, Elliott Sober, and Roger White, for helpful comments by three anonymous referees, and for audience comments on an earlier version of the paper presented at a conference by the London School of Advanced Study on the Doomsday argument (London, Nov. 6, 1998). I gratefully acknowledge a research grant from the John Templeton Foundation.

References

Bartha, P. and C. Hitchcock (1999). "No One Knows the Date of the Hour: An Unorthodox Application of Rev. Bayes's Theorem." *Philosophy of Science* (Proceedings) 66: S229-S353.

Bartha, P. and Hitchcock, C. (2000). "The Shooting-Room Paradox and Conditionalizing on Measurably Challenged Sets. " *Synthese* 108 (3): 403-437.

Bostrom, N. (1999). "The Doomsday Argument is Alive and Kicking." *Mind* 108 (431): 539-550.

Bostrom, N. (2000a). "Observer-relative chances in anthropic reasoning?" *Erkenntnis* 52: 93-108.

Bostrom, N. (2000b). "Observational Selection Effects and Probability." *Doctoral dissertation*, Department of Philosophy, London School of Economics, London. Available at http://www.anthropic-principle.com/phd/.

Bostrom, N. (2001a). "Are Cosmological Theories Compatible with All Possible Evidence? A Missing Methodological Link." In preparation. Preprint at http://www.anthropic-principle.com/preprints.html

Bostrom, N. (2001b) "A Super-Newcomb Problem." *Analysis*. In press.

Carter, B. (1983). "The anthropic principle and its implications for biological evolution." *Phil. Trans. R. Soc*. A 310: 347-363.

Carter, B. (1989). "The anthropic selection principle and the ultra-Darwinian synthesis." In *The Anthropic Principle*, eds. F. Bertola and U. Curi., Cambridge University Press, Cambridge, pp. 33-63.

Coles, P. & Ellis, G. (1994). "The Case for an Open Universe." *Nature*. Vol. 370, No. 6491: 609-615.

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

Elga, A. (2000). "Self-locating Belief and the Sleeping Beauty problem." *Analysis* 60.2: 143-147.

Freedman, W. L. (2000). "The Hubble constant and the expansion age of the Universe." *Physics Letters*. Vol. 333: (1-6): 13-31.

Gott, R. J. (1993). "Implications of the Copernican principle for our future prospects." *Nature* 363 (27 May): 315-319.

Gott, R. J. (1994). "Future prospects discussed." *Nature* 368 (10 March): 108.

Gott, R. J. (1996). "Clusters, Lensing, and the Future of the Universe." *Astronomical Society of the Pacific Conference Series*. Vol. 88, San Francisco, eds. V. Trimble and A. Reisenegger.

Grove, A. J. (1997). "On the Expected Value of Games with Absentmindedness." *Games and Economic Behaviour*. 20: 51-65.

Hall, N. (1994). "Correcting the Guide to Objective Chance." *Mind* 103 (412): 505-517.

Hawking, S. and Israel, W. (1979). *General Relativity: An Einstein Centenary Survey*. Cambridge, Cambridge University Press.

Kopf, T., P. Krtous, et al. (1994). "Too soon for doom gloom." Preprint at http://xxx.lanl.gov/abs/gr-qc/9407002.

Lachièze-Rey, M. and Luminet, J-P. (1995). "Cosmic Topology. " *Physics Report*. Vol. 254, no. 3: 135-214.

Leslie, J. (1989). *Universes*. London, Routledge.

Leslie, J. (1992). "Doomsday Revisited." *Philosophical Quarterly* 42 (166): 85-87.

Leslie, J. (1993). "Doom and Probabilities." *Mind* 102 (407): 489-91.

Leslie, J. (1996). *The End of the World: the science and ethics of human extinction*. London, Routledge.

Lewis, D. (1980). "A Subjectivist Guide to Objective Chance", in Richard C. Jeffrey, ed., *Studies in Inductive Logic and Probability*, vol. II. Berkeley: University of California Press. Reprinted with postscripts in Lewis 1986, pp. 83-132.

Lewis, D. (1986). *Philosophical Papers*. New York, Oxford University Press.

Lewis, D. (1994). "Humean Supervenience Debugged." *Mind* 103 (412): 473-490.

Linde, A. (1990). *Inflation and Quantum Cosmology*. San Diego, Academic Press.

Martin, J. L. (1995). *General Relativity*. 3rd edition, London, Prentice Hall.

Mellot, D. H. (1971). *The Matter of Chance*. Cambridge: Cambridge University Press.

Neta, A. and Bahcall, N. et al. "The Cosmic Triangle: Revealing the State of the Universe." *Science Magazine*. Vol. 284, no. 5419, Issue of 28 May 1999, pp. 1481-1488.

Nielson, H. B. (1989). "Random dynamics and relations between the number of fermion generations and the fine structure constants." *Acta Physica Polonica* B20: 427-468.

Olum, K. (2000). "The doomsday argument and the number of possible observers." Preprint at http://xxx.lanl.gov/abs/gr-qc/0009081.

Perlmutter,S. et al. (1999). "Measurements of Omega and Lambda from 42 high-redshift supernovae." *Astrophysical Journal* 517: 565-586.

Perry, J. (1977). "Frege on Demonstratives." *Philosophical Review* 86: 474-97.

Piccione, M. and A. Rubinstein (1997). "On the Interpretation of Decision Problems with Imperfect Recall." *Games and Economic Behaviour* 20: 3-24.

Reiss, A. (2000). "The Case for an Accelerating Universe from Supernovae." *Publications of the Astronomical Society of the Pacific* 122: 1284-1299.

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

Zehavi, I. and Dekel, A. (1999). "Evidence for a positive cosmological constant from flows of galaxies and distant supernovae." *Nature* 401 (6750): 252-254.

[1] I have elsewhere argued that if there are many extraterrestrial civilizations then the Doomsday argument doesn't work even on its own terms (Bostrom 1999), because the greater likelihood of you being a human (rather than an extraterrestrial) given a more populous human species compensates for the probability shift in favor of a less populous human species entailed by the Doomsday argument. Incidentally, the assumption that we are alone in the universe is almost certainly false if recent evidence suggesting we are living in an open universe is to be trusted. An open (or flat) universe, assuming the simplest (i.e. singly connected) topology, is spatially infinite and contains infinitely many stars and planets, and hence presumably infinitely many intelligent species. (More on this below.) So from an empirical point of view we are making a big concession when granting this assumption.

[2] We pretend that these are the only possibilities. Although it is trivial in principle to extend the argument to the more realistic case where a much larger set of hypotheses are given non-zero prior probabilities, in practice this would require some labor. In order to get maximal precision, one would have to assign prior subjective probabilities, based on all one's non-indexical empirical information, to the full range of

possible sizes of the human population, and Bayesian updating would have to be applied to each of these hypotheses separately. Richard Gott, one of the independent discoverers of the Doomsday argument, proposes (e.g. in (Gott 1996), in the context of a discussion of Jeffreys's Tram problem) the adoption of the prior $\Pr(N)dN \propto dN/N$ as a suitable vague prior for the doomsayer. This entails a 50% posterior probability of you being among the first 50% of all humans, a 10% probability of you being among the first 10%, and so on, making calculations trivial. However, this result rests on the specific empirical "improper" prior that was assumed, one that one need not accept. For example, one could argue (as does e.g. Leslie 1996) that *if* humankind survives long enough to begin to colonize space, then it will likely survive for a very long time and in very large numbers. Thus, the idealization considered in the text may not be too far off the mark.

[3] See e.g. Leslie 1992; Gott 1993; Leslie 1993; Gott 1994; Leslie 1996; Bostrom 1999. For an early version of the Doomsday argument, see also Nielson (1989). Strictly speaking, what the Doomsday argument purports to show is that the probability that there will be many more humans has been overestimated. This does not imply impending doom. The conclusion is compatible with the human species surviving for a very long time if population size declines sufficiently (which arguably, however, may constitute a type of doomsday scenario). Another possibility is that we evolve, or design ourselves, into some kind of beings who don't count as members of the reference class of observers used in the Doomsday argument. Moreover, John Leslie thinks that the Doomsday argument is substantially weakened if the world is indeterministic, although other doomsayers disagree with him on that point. As explained later, we shall bracket all these complications by considering only possible worlds where they do not pertain. We can then focus more sharply on the relevant philosophical issues.

[4] For the cognoscenti, it should be said that we are also assuming that the Self-Indication Assumption (which states, roughly, that finding that you exist gives you reason to think that there are many observers) is false. The Self-Indication Assumption has been embraced by some critics as a means of defeating the Doomsday argument (Dieks 1992; Kopf, Krtous et al. 1994; Bartha and Hitchcock 1999; Bartha and Hitchcock 2000; Olum 2000). An argument against the Self-Indication Assumption is given below in note 7.

[5] We suppose the incubator to be a mindless automaton which doesn't count as an observer.

[6] $\Pr(\mathit{Tails} \mid \mathit{I\ am\ in\ cell\ \#1})$

$$= \frac{\Pr(\mathit{I\ am\ in\ cell\ \#1} \mid \mathit{Tails})\Pr(\mathit{Tails})}{\Pr(\mathit{I\ am\ in\ cell\ \#1} \mid \mathit{Tails})\Pr(\mathit{Tails}) + \Pr(\mathit{I\ am\ in\ cell\ \#1} \mid \mathit{Heads})\Pr(\mathit{Heads})}$$

$$= \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{100} \times \frac{1}{2}} = \frac{100}{101}.$$

[7] This assumes that we reject the Self-Indication Assumption (SIA). If we instead accept SIA then at stage (a) you should be fairly confident that the coin fell heads, on grounds that on that hypothesis there would be more observers and thus a greater probability that you should find yourself having come into existence. If we explicate this principle to mean that hypotheses on which 2N observers exist give (other things equal) twice the probability to you finding yourself alive as do hypotheses on which merely N observers exist, then at stage (a) your credence in Heads should be 100/101. Continuing the calculation as in the main text, this leads to a posterior probability of Heads equal to 1/2. While SIA has certain attractive features (in particular that it cancels the Doomsday argument and the other strange results we shall get in later sections of this paper), it comes with a hefty price tag as shown in 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 to you that T_2 is about a trillion times more likely to be true than T_1 (whereupon the philosopher runs the *Incubator* thought experiment and appeals to SIA)!"

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 SIA in *Incubator* either.

[8] Such conditional probabilities can be nontrivial even if the right-hand side specifies exactly which possible world is the actual one; for there can remain uncertainty as to which position in this world I occupy – to use Quine's terminology, which *centered* possible world I am in. The metaphysics of indexical facts is not our concern here, but the point that one can learn something new when one discovers which person one is even if one already knew every non-indexical fact can be made via the story of the amnesiacs in the Stanford library (adapted from John Perry (1977); see also (Lewis 1986)): Two amnesiacs are lost in

the library on the first and second floor, respectively. From reading the books they have learned precisely which possible world is actual – in particular they know that two amnesiacs are lost in the Stanford library. Nonetheless, when one of the amnesiacs sees a map on the wall saying "YOU ARE HERE", with an arrow pointing to the second floor, he learns something he didn't know: that he is the amnesiac on the second floor.

[9] Every experience that a human could have is, it seems, probably had by somebody somewhere. This follows if we assume that cosmos is sufficiently big and that it contains a suitable class of physically random phenomena. In the actual world, it seems that we have many such phenomena: thermal fluctuation, black hole evaporation ("Hawking radiation"), and other types of quantum jitter. Because of such randomness, each finite chunk of spacetime, such as a galaxy or a black hole, has a finite probability of generating any modest-sized structured lump of matter such as a human brain in a particular state (Hawking and Israel 1979, p. 19). There is recent evidence suggesting that our universe is open or flat (e.g. Coles and Ellis 1994, Freedman 2000) and therefore, assuming it is singly connected, spatially infinite at every point in time (e.g. Martin 1995; for an introduction to singly versus mutiply connected topologies, see Lachièze-Rey and Luminet 1995). According to even more recent data, we seem to be living in a universe with a positive cosmological constant λ (Perlmutter et al. 1999, Zehavi, I. and Dekel 1999, Bahcall et al. 1999, Reiss 2000), which leads to an infinite universe in most plausible models that have been proposed. There is also the possibility that there are many other physically real universes beside our own (which is a consequence of the currently most popular versions of inflationary cosmology – see e.g. Linde 1990), which adds to the case for thinking that there is infinitely much stuff out there.

Given that the number of galaxies or black holes is infinite (or is finite but sufficiently large), it therefore follows that with a high probability – probability one (or infinitesimally close to one) in the infinite case – every possible brain state (of finite complexity) is instantiated somewhere. The thesis then follows, assuming that mental states supervene on brain states. (Sizable chunks of environment will also exist in all possible configurations somewhere, and so will brains that have evolved and are making veridical observations – for instance of measurement apparatuses that have also spontaneously materialized from random phenomena next to the observing brain.)

From a philosophical point of view, of course, these empirical assumptions are not crucial. Even if the number of observers in the world is in fact quite small, one may still maintain that our methodological toolkit ought to contain the resources needed to evaluate hypotheses according to which the world is big and random in the ways described.

[10] It is easy to show that if $\Pr(E|T_1) = 1$ and $\Pr(E|T_2) = 1$, then $\frac{\Pr(T_1 \mid E)}{\Pr(T_2 \mid E)} = \frac{\Pr(T_1)}{\Pr(T_2)}$.

[11] For a more detailed argument for this claim and an exploration of some of its consequences, see Bostrom 2001a. There remains the problem of how to deal with the case where the number of observers is not just very large but strictly infinite (and hence the relevant fraction is not a well-defined quantity in standard analysis). One may possibly try to approach this issue by using infinitesimal probabilities or alternatively by considering densities of observers.

[12] We assume that Eve and Adam and whatever descendants they have are the only inhabitants of this world. If we assume, as the Biblical language suggests, that they were placed in this situation and given the knowledge they have by God, we should therefore also assume that God doesn't count as an "observer" in the relevant sense here. Note that for the reasoning to work, Adam and Eve must be extremely confident that if they have a child they will in fact spawn a huge species. One could modify the story so as to weaken this requirement, but empirical plausibility is not an objective in this thought experiment.

[13] John Leslie does not accept this result and thinks that Eve should not regard the risk of pregnancy as negligible in these circumstances, on the grounds that the world is indeterministic and the SSA-based reasoning runs smoothly only if the world is deterministic or at least the relevant parts of the future are already "as good as determined" (personal communication; compare also Leslie 1996, pp. 255-6, where Leslie discusses a somewhat similar example). I disagree with his view that the question about determinism is relevant to the applicability of SSA. But in any case, we can legitimately evaluate the plausibility of SSA (with an unrestricted reference class) by considering what it *would* entail if we knew that the world were deterministic.

[14] Note that if he intends to repeat the experiment then the number of offspring that he would have to intend to create increases. If the prior probability of the outcome of a deer appearing is one in ten thousand and the trials are independent, then if he wants to do the experiment twice he would have to intend to create at least on the order of ten million offspring. If he wants to repeat it ten times he would have to intend to create about 10^{40} offspring to get the odds work out in his favor.

[15] The reason why there is a discrepancy between what Adam should believe and what the external observer should believe is of course that they have different information. If they had the same information they could agree (Bostrom 2000a).

[16] The parts of Lewis's theory that are relevant to the discussion here can be found in chapters 19 and 21 of (Lewis 1986).

[17] I'm simplifying in some ways, for instance by disregarding certain features of Lewis' analysis designed to deal with cases where there is no closest possible world, but perhaps an infinite sequence of possible

worlds, each closer to the actual world than the preceding ones in the sequence. This and other complications are not relevant to the present discussion.

[18] If he did know that we exist, then it would definitely *not* be the case that he should give a high conditional probability to *C* given *E*! Quite the opposite: he would have to set that conditional probability equal to zero. This is easy to see: By the definition of the thought experiment, we are here only if Adam has a child. Also by stipulation, Adam has a child only if either doesn't form the intention or he does and no deer turns up. It follows that if he forms the intention and we are here, then no deer turns up. So in this case, his beliefs would coincide with ours; we too know that if he has in fact formed the intentions then no deer turned up.

[19] Under the supposition that if there is *AC* then there is *C*, the hypothesis that there will be *C* conflicts, of course, with our best current physical theories, which entail that the population policies of UN^{++} have no significant causal influence on distant gamma ray burst. However, a sufficiently strong probability shift (resulting from applying SSA to the hypothesis that UN^{++} will create a sufficiently enormous number of observers if *C* doesn't happen) would reverse any prior degree of confidence in current physics (so long as we assign it a credence of less than unity).

[20] The reason this question doesn't seem relevant to the evaluation of SSA is that the answer is likely to be "spoils to the victor": proponents of SSA will say that whatever SSA implies is rational, and its critics may dispute this. Both would be guilty of question-begging if they tried to use it as an argument for or against SSA.

[21] I discuss a related decision problem, the "Super-Newcomb Problem", in (Bostrom 2001b).

[22] In fact, if we accept SSA we should think this situation astronomically unlikely – about as unlikely as the coincidences would be! We can see this without going into details. If we ever get to the situation where UN^{++} executes the plan then one out of two things must happen, both of which have extremely low prior probabilities: a series of strange coincidences, or – which is even more unlikely given SSA – we happen to be among the very first few out of an astronomically large number of humans. If P_1 implies that either P_2 or P_3, and we assign very low probability both to P_2 and to P_3, then we must assign a low probability to P_1 as well.

[23] I explore one way of modifying SSA (by relativizing the reference class) in chapter 9 of (Bostrom 2000b). Such a move could break the chain of reasoning that leads to the weird conclusions discussed above. The difficulty is to make sure that the relativized principle supports all the legitimate uses of SSA, including the thought experiments and the considerations from the methodology of cosmology used to motivate its introduction in this paper. It is still an open question whether this can be done.