© F **Philosophical Writings**
No. 27 Autumn 2004

AGENCIES, CAPACITIES, AND ANTHROPIC SELF-SELECTION

Milan M. Cirkovic

Abstract

Several intriguing thought experiments have been recently devised by Bostrom (2000, 2001) in order to show that the so-called Self-Sampling Assumption (SSA) leading to the (in)famous Doomsday Argument (DA) needs further specification in order to avoid an array of seemingly paradoxical and unnatural consequences. These consequences have been used by proponents of the rival Self-Indication Assumption as a indication of general insufficiency of SSA and, consequently, the fallaciousness of the Doomsday Argument. Bostrom has also constructed a method of judging counterfactuals in order to avoid the apparent incoherencies thus entailed. Here we would like to point out that there is a sort of cheaper (in epistemological terms) way out of the difficulties, dealing with limited capacities of the agencies considered. That this sort of way out is not obvious represents another instance of the notorious coherence gap problem in thought experiments. The similarity of the situation in the field of anthropic self-selection with the one in the field of time travel and backward causation is briefly considered.

1. Introduction: Adam, Eve and all that

In a recent inspiring and thought-provoking study, as well as in some earlier writings, Nick Bostrom (2001; see also Bostrom 1999, 2000) has analyzed consequences of a general assumption usable in the theory of observation selection effects, dubbed the Self-Sampling Assumption (henceforth SSA) and defined in the following way:

> *The Self-Sampling Assumption.* Every observer should reason as if she were a random sample drawn from the set of all observers.

SSA is a methodological prescription stating how reasonable epistemic agents should assign credence and make probabilistic inferences in situations involving observational selection. We can see its operation in Bostrom's thought experiments discussed in detail below, as well as in its most notorious consequence – the so-called Doomsday Argument (henceforth DA; Gott 1993; Leslie 1996, and extensive bibliography therein).[1] Numerous opponents of the DA conclusion have, naturally, turned their analytical artillery toward SSA. On the other side, there is a strong motivation for investigating all consequences of SSA for those who, like Bostrom (2000), accept its general validity (provided that the reference class is relativized in some way). In the course of his detailed analysis, Bostrom (2001, 2002) lists several seemingly surprising and paradoxical consequences of SSA illustrated by beautiful thought experiments; these consequences include backward causation, paranormal causation and psychokinesis. We shall quote three out of four such experiments described by Bostrom here in full in order to make the arguments in next section easier to follow. The remaining experiment (also involving Adam and Eve) is conceptually similar to the second one, and therefore is omitted here.

> *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. 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, on the other hand, Eve doesn't

[1] The core idea of DA can be expressed through the following urn-ball experiment. Place two large urns in front of you, one of which you know contains ten balls, the other a million, but you do not know which is which. The balls in each urn are numbered 1, 2, 3, 4, ... Now take one ball at random from the left urn; it shows the number 7. This clearly is a strong indication that the left urn contains only ten balls. If the odds originally were 50:50 (identically-looking urns), an application of Bayes' theorem gives the posterior probability that the left urn is the one with only ten balls as P_{post} (n=10) = 0.99999. Now consider the case where instead of two urns you have two possible models of humanity, and instead of balls you have human individuals, ranked according to birth order. One model suggests that the human race will soon become extinct (or at least that the number of individuals will be greatly reduced), and as a consequence the total number of humans that ever will have existed is about 100 billion. The other model indicates that humans will colonize other planets, spread through the Galaxy, and continue to exist for many future millennia; we consequently can take the number of humans in this model to be of the order of, say, 10^{18}. As a matter of fact, you happen to find that your rank is about sixty billion. According to Carter and Leslie, we should reason in the same way as we did with the urn balls (that is, apply SSA). That you should have a rank of sixty billion is much more likely if only 100 billion humans ever will have lived than if the number was 10^{18}. Therefore, by Bayes' theorem, you should update your beliefs about mankind's prospects and realize that an impending doomsday is much more probable than you thought previously.

> 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!"
>
> *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.
>
> *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^{++} rises 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*.

In order to see how the strange consequences follow from SSA, let us briefly consider the first experiment in detail. We may set the prior probability that Eve conceive a child as about 0.5, and let us suppose that there will actually be 100 billion people constituting Adam's and Eve's offspring as a consequence of such event (very roughly corresponding to the actual number of humans having existed to this day). Probability of having a birth rank less or equal to 2—under the assumption of only two human beings existing is—of course, 1. The use of SSA (and this is a crucial point) gives that the probability of having a birth rank less or equal to 2 under the assumption of 10^{11} humans existing is about 5×10^{-11}. Using Bostrom's notation, we may denote birth rank by R and the total number of humans existing as N. According to the theorem of the Rev. Bayes, the posterior probability of

$$\Pr(N = 10^{11} \mid R \le 2) = \frac{\Pr(R \le 2 \mid N = 10^{11})\Pr(N = 10^{11})}{\Pr(R \le 2 \mid N = 10^{11})\Pr(N = 10^{11}) + \Pr(R \le 2 \mid N = 2)\Pr(N = 2)}$$
$$\approx \frac{5\times 10^{-11}\cdot 0.5}{5\times 10^{-11}\cdot 0.5 + 1\cdot 0.5} \approx 5\times 10^{-11}.$$

Thus, risks of getting pregnant are indeed negligible, and a form of anomalous causation seems to occur. Similar analyses apply to other thought experiments as well, and are presented in a vivid style by Bostrom (2001, 2002). In particular, it is claimed by the same author that the UN^{++} thought experiment described a situation which may become a realistic possibility at some time in future.

We are, therefore, led to *prima facie* believe that backward and paranormal causation (on literally cosmic scales!), as well as psychokinesis are built in the anthropic reasoning leading to SSA and the infamous Doomsday argument. (Here and elsewhere, we are using the sensitive term "anthropic" in its uncontroversial meaning of "pertaining to the observational selection effects", without presupposing any particular explanation of such effects.) As a solution to such counterintuitive consequences of SSA, Bostrom (2001) suggests the following mechanism. In the world of Adam and Eve, Adam is justified in using SSA, but he will not, if actually performing an experiment, witness the anomalous coincidence, since he does not possess the same information as we do as outside observers ("comparing his world to ours"). Although SSA is thus misleading from the Adam's standpoint, it is not necessarily (or even *prima facie*) incorrect, because the theory of counterfactuals employed by Bostrom uses a subtle loophole: if action A is believed to bring about C, the conjuction of ¬A and the statement "it is false to assume that have A occurred, C would have occurred" is coherent. In this manner, one correctly accounts for **our** not expecting anomalous causation to occur. However, at least a part of the price to be paid for such a description is the necessity to fix a specific time *t* at which we **must** decide whether A occured or not; since the simultaneity of Adam's action and our perception of whether he performed it or not is observer-relative, this situation seems to entail temporal becoming and therefore the description necessarily subscribes to A-theories of time. However, A-theories do not square well with the modern physical concepts (e.g. Grünbaum 1973). In addition, Bostrom's description seems to entail indeterminism (for instance, when claiming that "whether there is a coincidence or not in a world presumably makes little difference as to how similar it can be to *w* [the actual world in the sense of David Lewis] with respect to its history up to *t*") which, although certainly an open issue, has the unappealing side of involving us in much wider controversy.

While accepting the main conclusion of the Bostrom (2001) discourse—the applicability and usefulness of SSA in the field of anthropic self-selection—we shall attempt to show here that his appeal to the theory of counterfactuals is unnecessary, and even slightly confuses the main issue. It is indicative that in a recent study rejecting SSA, Ken Olum (2002) explicitly cites the "paradoxes" discussed here as arguments for the fallaciousness of SSA and DA, and Bostrom's explanation qualified as "some rather strange argumentation".

In what follows, we shall attempt to show that the simplest way out of the "paradoxes" of SSA entails re-assessment of the possibility and properties of the entities postulated in these thought experiments. In particular, we shall show that the wide **coherence gap** arising in these situations allows us to deny the coherent existence of such agencies and their assumed capacities. Several similar situations encountered in the philosophy of space and time, as well as the philosophy of religion, are briefly discussed and the (rather well-known and simple, although admittedly sometimes unrecognized) common solution to the problems indicated. Finally, we shall try to show that this solution, while seemingly counterintuitive, is itself coherent and leaves SSA and related assumptions unscathed.

2. Agencies and their capacities

The issue of capacities of various agencies (especially extraordinary or "exotic" ones) has a long and colorful philosophical and theological history. In XI century of the Christian era, Peter Damian defended the idea that God's omnipotence extends so far that He can change the course of past events. In a memorable passage of his most famous tractatus *De Omnipotentia Dei*, Damian wrote that

> ...just as we can duly say 'God was able to make it so that Rome, before it had been founded, should not have been founded,' so in the same way we can equally and suitably say, 'God can make it so that Rome, even after it was founded, should not have been founded'... If therefore it is coeternal with God to have power over all things, then God can make it so that those things which were done shall not have been done. But it is coeternal with God to have power over all things. Therefore God can make it so that what has been done shall not have been done.[2]

After a prolonged discussion (which is of some interest to this day; see, for instance a clear and refreshing treatment in Remnant 1978), most theologians, including Thomas Aquinas and St. Bonaventure, concluded that Damian was wrong in attributing inherently paradoxical capacities to the deity. However, alternative explanations are possible (cf. McArthur and Slattery 1974), invoking non-reality of past facts. This is of interest to our present subject not only because Adam and Eve experiments are located in a classical setting necessitating the presence of the same deity, but also because it **(i)** illustrates a caution necessary in consideration of any agency different from those encountered in our experience, and **(ii)** represents a peculiar form of backward causation. As Remnant notices:

> It seems to follow that if we are unable to change the past but are able to change the present and the future, then our inability does not result from it being *logically* impossible to change the past, but from some unique feature of our relationship, in the order of nature, to past events, as contrasted with our relationship to present and future events.

With this in mind, let us now consider for a moment one of the greatest problems of physics and philosophy of all times: the issue of time travel. As recognized long ago, the major conceptual problem with the time travel entails the issue of backward causation, i.e. the possibility of time-travelling agent changing the past in such a manner that a paradoxical state-of-affairs is created in which both events A and ¬A occur. This is traditionally dubbed the bilking paradox.[3]

[2] Quoted after the translation of McArthur and Slattery (1974).

[3] This sort of paradox arises when one claims that it is possible to bring about an earlier event A through a later event B. To see paradoxical consequences of such correlation

Now we need to distinguish two related but different issues. In some of the SSA-related paradoxes, the issue of backward causation is also involved, apart from the problem of paranormal causation we are primarily dealing here with. In particular, Bostrom's experiment 4 and related variations (even if a nearby γ-ray burst has or has not already occured in the past, in view of the finite speed of information propagation, we can change the epistemic probability[4] of its having occured by intentionally modifying the number of our offspring in order for us to achieve the desired rank in the entire human population; see Olum 2002) undoubtedly entail a sort of backward causation. However, we do not find this to be a problem for SSA, at least not in the sense and measure as it is for time travel, since no possibility of bilking appears in the former case, as will be discussed elsewhere (Cirkovic, manuscript in preparation).

Even if not of immediate concern for SSA, the lively philosophical debate on time travel and backward causation has very important lessons to teach those engaged in anthropic thinking. The most important one is the lesson on difference between our common-sense capacities and extraordinary capacities which are prone to appear in various thought experiments concerning time travel. To cite a related example, Paul Tappenden (2000) writes:

> We do not currently know of any reason why causal loops are impossible. Setting quantum mechanics aside, there is the famous 'grandmother paradox', but that is easily resolved. If I were to go back in time I could not kill my grandmother before she gave birth to my mother however hard I tried. Necessarily, something would always frustrate my efforts.

The nature of the frustrating agent is not obvious, but the author seems rather certain that frustration will occur in the prescribed situation.

between future and past, it is enough to introduce an additional event C which occurs as a consequence of A and which **prevents** B from occuring. In already conventional terms, the correlation between A and B is bilked by C (Flew 1954; Mellor 1981). This problem lies in the background of all stories about a hero travelling to the past and killing one of his ancestors, which prevents him from existing and travelling to the past, etc.

[4] However, if the series of coincidences that we should think would happen were construed as a causal relation (say by establishing a long, statistically significant, chain of coincidences), then it would be justified to say that our intuitions would actually have changed the chance.

Commenting upon some operationalizations of the same old bilking paradox appearing in the literature, Bryson Brown (1992) emphasizes:

> In both Davies's and Mellor's arguments the combination of backwards causation with the exercise of capacities we ordinarily take ourselves to have is what leads to trouble. But capacities and their relation to physics already comprise a difficult and controversial subject. So we must explore in more detail the assumptions about capacities that are required to make these arguments against time travel work, and see if there are any reasonable alternatives to them.

The same lesson—as well as the bulk of the Brown's ingenious discussion—may be applied to the analysis of SSA-related paradoxes. In all Bostrom's thought experiments, the crucial point is that the subjects display willpower of volition; for instance, UN^{++} needs to make a resolution or some such proclamation establishing a programme of building space colonies. However, we all know from practical experience that it frequently occurs that the human resolutions proclaimed with absolute sincerity and with much smaller requirements (when resources and time are concerned) remain unfulfilled. The same applies to the Adam's and Eve's decision to have or have not children. This may be interpreted to mean that although such intentions are sincerely proclaimed, they are not true capacities of the subjects, and the required correlations will fail. It may, however, still be stated with certainty that **if the relevant subjects (Adam, Eve, UN^{++}, etc.) possessed such capacities, they should think that this would lead to paradoxical consequences**, in particular precognition and paranormal causation. This is the standard compatibilist interpretation.

Other analyses in the voluminous literature on time travel apply to this issue. For instance, John Earman (1995, p. 171-172) points out to the true cause of concern:

> Suppose that Kurt tries over and over again to kill his grandfather. Of course, each time Kurt fails—simetimes because his desire to pull the trigger evaporates before the opportune moment, sometimes because although his murderous desire remains unabated his hand cramps before he can pull the trigger, sometimes because although he pulls the trigger the gun misfires, sometimes because although the gun fires the bullet is deflected, etc. In each instance we can give a deterministic explanation of the failure. But the obtainment of all the initial conditions that result in the accumulated failures may seem to involve a coincidence that is

> monstrously improbable... Here we have reached a real issue but one which is not easy to tackle.

But in the cases of SSA-produced causal anomalies, it is exactly this statistical plank which is missing! Adam and Eve cannot conceivably repeat their presumed exercises of anomalous causal powers in order that bizarre coincidences appear above the noise level (created by "truly" random wandering limping deer). Yet less is the same kind of repetitions possible for UN^{++} in the prescribed situation.

On the view of compatibilism, although Adam and Eve and UN^{++} have capacity to perform actions which would cause unlikely correlations, they do not, and so no contradictions or paradoxical consequences arise. Therefore, the solution lies in making the difference between paradoxical and merely counterfactual actions. Reliance on our intuitions in regard of capacities of the first humans or members of the future all-poweful government is simply mistaken.[5] Both Adam and Eve, and members of UN^{++} differ sufficiently from us in a way which justify denying them capacities we take ourselves to have in similar circumstances (or just take for granted).

We may reach the same conclusion from a slightly different starting point, and use the opportunity to put the problem in a wider epistemological picture. The general (and unfortunately rather rarely explicitly treated) difficulty arising in all discussions of thought experiments is the problem of the coherence gap. Thus Ivan Havel (1999) writes:

> In conceivable worlds of thought experiments, some states-of-affairs are, by design, the *same* as they are in our world, while other states-of-affairs are deliberately *different*... The crucial but often neglected feature of these worlds is that we seldom know what is the extent of the domain of "the same" and what is the extent of the domain of "the different", besides what is explicitly mentioned or used in the construction. Moreover, besides these two domains there is an inexhaustible realm of states-of-affairs that are *omitted* because they are believed to be irrelevant or because they are forgotten, obscured or entirely beyond the reach of human knowledge.

[5] An illustrative parallel to UN^{++} can be found in dystopias dealing with similar all-powerful bodies. For instance, in the literary context of Orwell's *1984* it slowly and painfully becomes plausible from the viewpoint of the main character that capacities of Big Brother and the highest Party circles are different from the common-sensical human ones.

Omitted realm of states-of-affairs constitutes the **coherence gap**, and the question of its possible impact on our reasoning the **coherence gap problem**.[6] We need to be highly cautious in evaluating any thought experiment not because it may entail empirical difficulties, but because the conceived world of the particular thought experiment may need additional assumptions or constraints in order for the desired outcome to occur. The schematic representation of the coherence gap is shown in Figure 1.

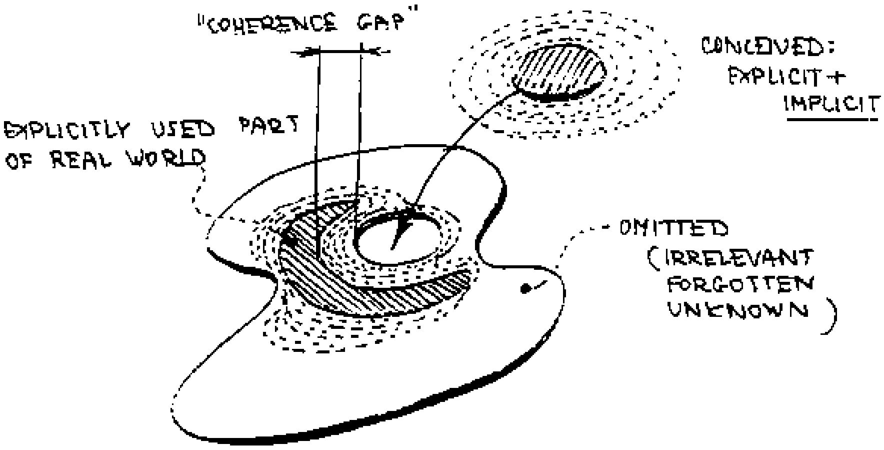
"COHERENCE GAP"
CONCEIVED: EXPLICIT + IMPLICIT
EXPLICITLY USED PART OF REAL WORLD
OMITTED (IRRELEVANT FORGOTTEN UNKNOWN)

Figure 1. The coherence gap problem as sketched by Havel (1999).

Width of the coherence gap varies depending on each particular scenario, but the general tendency is intuitively clear: the farther the conceived situation lies from the real world (and it should be assumed, as in Havel's study, that we are methodological realists), the wider the coherence gap is. To use Havel's subtle example, the famous Newton's bucket experiment is—**when explicitly treated as a thought experiment**—very "realistic" and its coherence gap is rather narrow. (Incidentally, as the said experiment in conjunction with the entire controversy surrounding Mach's principle shows, *new* physical ideas are often hidden in coherence gaps of *old* and seemingly completely understood thought experiments.) In our case, we may cite Bostrom himself on his UN^{++} thought experiment:

[6] Havel's paper is interesting from another point of view related to the anthropic reasoning: it explicitly treats the "implied observer" problem in conceived worlds. However, we cannot enter into that discussion here.

> The main new feature of this experiment is that it depicts a situation that we can potentially actually bring about. Creating UN^{++} might be practically difficult, and there is no guarantee that other preconditions are satisfied (that there are no extraterrestrials, for example); yet it is the sort of undertaking that could quite conceivably be accomplished through conventional non-magical means.

The crucial word in the Bostrom's account is **conceivably**. Again, as emphasized by Havel, the notion of conceivability is wider than the notion of possibility; a conceived world may not be possible world, even if superficially coherent, due to the coherence gap problem. What we claim here is that **UN^{++} with the attributed capacities is conceivable, but not possible**. There may be various concrete reasons for that, some of which Bostrom correctly mentions. There is, however, one even more relevant for the considered situation: the validity of SSA itself. Namely, if SSA is the correct assumption in probabilistic arguments concerning intelligent beings, we expect that conclusions of the Doomsday Argument are essentially valid. Therefore, it is highly unlikely that humans will continue to exist for a sufficiently long time in the future in order to create the UN^{++}.[7] In this manner, we may claim that the UN^{++} thought experiment (the one Bostrom states may be actually realized!) is actually incoherent. Paradoxical consequences follow from the application of an assumption (SSA), whose consequent application elsewhere actually prevents the main agency of the “paradox” to emerge.

Even if UN^{++} is ultimately formed, from the physical point of view, there is a host of possible reasons for its **incapacity** to fullfil the programme outlined in the thought experiment. For instance, there may not be enough resources available in the volume of space accessible to humans to create a sufficient number of space colonies. Or there may not

[7] Bostrom (2001) admits it in the footnote 22, claiming somewhat cryptically that “in fact, if we accept SSA we should think this situation [i.e. the one in which UN^{++} has the opportunity to execute its plan] astronomically unlikely – about as unlikely as the coincidences would be!” *Ultimo facie*, this may not be the exclusive explanation, since among ways to accept the DA conclusions are some “non-catastrophic” ones which would still make the realization of UN^{++} possible *in principle* (even as, say, the supreme executive body of posthumans, cyborgs, or anybody else falling outside of the reference class). Thus, we may suppose that UN^{++} decides—lacking the required capacity of creating humans in sufficiently large numbers—to populate new colonies with androids which, according to narrow reference class views, are not comparable to humans in the SSA sense (although they might be in some other, for instance ethical, sense). However, it seems inescapable that these other options are contingent upon solution of the reference class problem, a matter we cannot pursue here further.

be enough time to do that. It is even conceivable that some yet unknown preemptive physical mechanism exists, similar to the mechanism involved in the solution of time travel paradoxes inspired by Wheeler-Feynman time-symmetric electrodynamics, as discussed by Cramer (1983), and ably described by Gregory Benford in his famous novel *Timescape* (Benford 1980). These and other similar obstacles are not just technical details, but various possible incoherencies lurking in the coherence gap.

These considerations may be summarized in the following way: the processes (political, sociological, neurophysiological, genetical, etc.) which **could** lead to the formation of UN^{++} are changing the entire setup of the thought experiment and modifying the intrinsic capacities of such a body, thus either resulting in no capacities at all (corresponding to non-formation of UN^{++}) or prescribing the incapacity to perform—when faced with the external situation as described—the plan Bostrom conceived.

As far as the Adam and Eve thought experiments are concerned, they are also prone to the coherence gap problem. In the first place, their particular conceived world contains an element which is not explicated in the thought experiments, but is highly relevant for any anthropic reasoning: the Divine creative agency. The presence of the deity causes severe problems for any thought experiment involving observers. Although it is not obvious, it is enough to use a more refined version of the same underlying idea, dubbed Strong Self-Sampling Assumption (SSSA) by Bostrom with very strong justification:

> ...We can take a first step towards specifying the sampling density by substituting *"observer-moments"* for "observers". Different observers may live differently long lives, be awake different amounts of time, spend different amounts of time engaging in anthropic reasoning etc. If we chop up the stretch of time an observer exists into discrete observer-moments then we have a natural way of weighing in these differences. We can redefine the reference class to consist of all observer-moments that will ever have existed. That is, we can upgrade SSA to something we can call the *Strong Self-Sampling Assumption*:
>
> *(SSSA)* Every observer at every moment should reason as if their present observer-moment were randomly sampled from the set of all observer-moments.

While we accept that SSSA is superior in many respects to the plain SSA, it raises a couple of issues in conceivable situations dealing with radically different observers from human ones ("exotic mentalities" in Bostrom's parlance). Unfortunately for the Adam-and-Eve sort of thought experiments, their situation is burdened by exactly such an "exotic mentality" problem, reflected in the necessary presence of the divine creative agency. Obviously, such (extremely) non-standard agency by its very presence (and definition) jeopardizes the consistency of the entire story. This is particularly visible if we stick to SSSA, since then we may rightly assume that the by far predominant statistical weight of Divine observer-moments indicate that practically every time we randomly choose an observer-moment it will be the God's one, and never Adam's or Eve's (or one of any of the present-day humans, for that matter). There are two principal possibilities: either the number of Divine observer-moments is infinite, or it is a very large finite number (obviously, it is not necessary to elaborate why we expect that number, even if finite, to be extremely large by human standards). In the infinite case, the reasoning above applies. For the very large finite case, the same applies, except that it is now just **very likely** that a randomly chosen observer-moment is Divine, and the simplest model will suggest that only once in $N_{\mathrm{Deity}}/(N_{\mathrm{Adam}} + N_{\mathrm{Eve}})$ times—where N_{being} denotes the number of observer-moments assigned to any one particular being—the conditions necessary for the experiment as Bostrom conceived it will be realized. (And even then, the serpent's probabilistic reasoning will be seriously weakened, since one may expect that N_{Deity} is large even compared to the total tally of all 10^{11} humans that will have ever existed.) Substituting God for an automaton—as Bostrom (2000) does—will not really help, since it is rational to conclude that the same capacity which requires creation of Adam and Eve (never mind the entire surrounding universe!) immediately implies an infinite number of observer-moments (in the Augustinian sense), or at least some very large such number. And any large number also spoils the point of the thought experiment. Therefore, if one wishes to have an automaton, then either this automaton will have an infinite or very large amount of observer-moments, and according to SSSA, any randomly chosen observer-moment will most likely be the automaton's, or—if one insists that automatons do not possess observer-moments by definition—it needs a sophisticated constructor, presumably so sophisticated that his/her tally of observer-moments is of similarly high order of magnitude as in the case of the personal deity of the Biblical myth.

Interestingly enough, this is the point of contact between two big problems in the field of anthropic self-selection: the problem of seemingly paradoxical consequences of SSA and the reference class problem. The latter can be roughly formulated in the form of a simple question: who counts as an observer? The definite solution of the reference class problem would entail the detailed definition of the capacities of members of a particular reference class in any given situation. This desired description may be very difficult to practically achieve, and that is why the reference class problem is so ubiquitous in anthropic reasoning and hard to solve. The very fact that the reference class problem remains unsolved (Bostrom 2000) warrants the approach in the present study, where we consider plausible differences in capacities of particular agencies in comparison to our everyday expectations.

In order to see in an example how hidden incoherencies may undermine Bostrom's account, let us (in connection with the "hunting with willpower" gedanken) consider the following conditional: *If there were only two human beings, the surface density of limping deer on Earth would be two orders of magnitude higher than in real forests*.[8] Such a state-of-affairs is unspecified in the original setup of the thought experiment, but we (as well as Adam and Eve!) may find it tentatively plausible and further specify the situation by postulating it, therefore reducing the size of the coherence gap. The truth value of such a statement is very hard to ascertain (although one may notice a slight reason for its plausibility in the fact that absence of a large population of very efficient human predators will sharply increase the deer population **in general**, so one may expect sharp increase in the absolute magnitude of the wounded fraction also). Now, Adam has a hundred times more reason to expect correlation to occur (in this **modified version** of the original thought experiment), although he—supposedly lacking the detailed knowledge of ecology (which, it is important to notice, **we are lacking too**)—may not realize that it has no causal connection with his decision. It seems inescapable that his knowledge or ignorance of the forest ecology changes his capacities in this version of the gedanken. Among other things, this shows that the full specification of capacities of an agency in a non-standard position (and consequent reducing of the coherence gap) is not an easy task.

Apart from inherently different capacities in comparison to the intuitive human ones, Adam's volition can hardly influence any one of

[8] We accept Havel's statement that the real world is coherent by definition. Therefore, there is a single well-defined (average) surface density of lame deer in real forests, which is presumably determined by complex ecological and physiological factors, which are still partially elusive to human science.

many known and probably even more unknown external parameters, such as free energy or spatial density constraints (let alone the possible existence of other observers, if we accept the no-outsider requirement). To see this clearly, let us compare two possible histories of humanity: one in which Adam and Eve carry out their plan, have a lot of children, and their children have children, etc. until the present population of humans, measurable in billions, is reached, and the other in which Adam and Eve have children, and their children have children, etc. for a couple of thousand years until a doomsday brought about by a stray asteroid or a pandemic of contagious disease. Of course, in the meantime, Adam and Eve have been exiled from Eden, became mortal, and died of old age, so they cannot distinguish which of the two histories is correct. Obviously, the argument motivating their decision has (if one accepts SSA) strength in the first case, but not in the second (or at least the Bayesian probability is drastically different for the same prior). Even if conscious of the physical obstacles—such as the possible extinction of humanity—Adam under no circumstances can influence them.

The UN^{++} example is even more tractable to this approach, since—as Bostrom correctly points out—it lacks the supernatural settings necessary for the Adam and Eve thought experiments. It is highly debatable whether, due to chaotic elements in human behaviour and nature, a government such as UN^{++} can be formed at all, even if no external obstacles were encountered. Another subtle issue lies in the possibility of the following situation. We may ascertain (by tremendous advances in sociological modeling, say) that UN^{++} is possible. However, even in that situation, we could never be certain that, after we form the universal government, we had managed to create UN^{++} rather than some other form of government which looked very similar to it, but lacked the required capacities to entail the anomalous causation. In other words, a proof of possibility is still very far from a proof of **inevitability** (on a compatibilist view). One of the plausible instatiations of such scenarios would be a situation in which, even if formed, such government would suppress human creativity and ingenuity (even purely technical ingenuity) to such a degree that creation of the required number of space colonies became impossible. Problems may also arise due to the finite future age of Earth, Solar System and other celestial bodies, which is subject of a new astrophysical discipline dubbed *physical eschatology* (Adams and Laughlin 1997; Cirkovic 2003). Further doubts concern the feasibility and efficiency of any form of production of human beings in sufficiently high numbers necessary for the alleged anomalous causation.[9] All these and other issues hang in the (rather wide) coherence gap of this colorful thought experiment.

3. Conclusion: no real threat to SSA

We conclude that there is a solution to the problems of seemingly paradoxical paranormal causation implied by SSA, different from the ones proposed by both its critics and defenders. While critics (like Olum) imply that paranormal causation illustrated by thought experiments like the ones described above indicates at least insufficiency (if not outright falsity) of SSA, defenders (like Bostrom) use a sophisticated and at least partially counter-intuitive reasoning based on not completely incontroversial foundations, like Lewis' theory of counterfactuals, in order to overcome the difficulties.[10] A much simpler hypothesis is the one proposed here, and that is the incompatibility of agencies such as postulated in the thought experiments with the situations described. We may accept the seemingly outlandish conclusions of the thought experiments while still questioning the relevant capacities of the agencies considered. This is not a case of suggesting empirical impossibility (although much can be said in favor of such, from a strictly empiricist point of view), but rather of pointing to the relevance of the coherence gap problem for the situations described. That part of the real world which needs to be substituted by the conceived situation may hide many different subtle incoherencies (as explained in the above analysis of the Adam and Eve thought experiments). This circumstance indicates that the burden of proof lies with those who would like to use these thought experiments as an argument against the usage of SSA (or SSSA or any other similar assumption) in the anthropic reasoning. Thinkers rejecting SSA for the alleged anomalous causation are thus expected to reduce the coherence gap by correctly specifying capacities of various agencies involved. In the absence of such a specification, the assertion of the implausibility of SSA has no real force.

If one feels the explanation unsatisfactory, an additional explanatory layer consists of analysis of causal (in)efficiency of conscious decisions. While allowing for the agencies involved to state

[9] The only method conceivable so far would be cloning (coupled with other advanced biomedical technologies), and both the practical efficiency and long-term consequences of cloning are subject to considerable debates at present.

[10] Although it should be mentioned that Bostrom claims (2001; private communication) that his conclusions are independent of the choice of the theory of counterfactuals.

their intentions, we deny the option of their decisions being carried out in the physical (even if only conceivable) world with any chance high enough to infer the required correlations damaging to our conventional notions of causality. It is rather well-known that the causal efficiency of even the Divine volition has been repeatedly questioned by philosophers criticising the theological approach to cosmology (e.g. Grünbaum 1998). How then can we expect the situation related to causal efficiency of decisions of humans and their institutions (like the UN^{++}) to be radically different? It is important to understand that the same medicine may be prescribed for the cases of backward causality; however, it is the opinion of the author that in these cases the medicine is too strong, since the backward causation in the anthropic context does not present any substantial physical problem, in contradistinction to the situation in the context of time travel. It does not lead to bilking paradoxes, as will be further explicated in the forthcoming study. Moreover, if B-theorists of time are correct, advanced physics entailing backward causation may be ubiquitous in the actual world (Price 1996), so that it may well turn to SSA's advantage!

One may find that the solution presented here is similar to Bostrom's account on the following points. When analyzing the Adam and Eve thought experiments, Bostrom describes a class of outside observers ("us") in possession of additional information, on the basis of which we conclude that the anomalous coincidence will not occur. In such a setup, Adam (as well as UN^{++} or any other entity we have qualified as non-standard agency) is necessarily in a different situation from "us". Our suggestion is that it automatically endows him with non-standard capacities. Bostrom does not say so explicitly, but stressing that "Eve and Adam were in highly unusual circumstances" comes to something very similar. It is important to notice that Bostrom's description does (although not very transparently) take into account physical variables of the system considered and their constraints; thus he mentions that "the presence of a correlation... would entail a world that would be somewhat different regarding the initial states of the deer". In a sense, both Bostrom's and the present account are two sides of the same coin, the present one being "physically-oriented" and Bostrom's rather "epistemologically-oriented" description.

The overall conclusion is that in any case SSA proponents (and, by extension, proponents of the Doomsday Argument) do not have anything to fear from the conceived situations involving exotic causal relations. Even if strikingly counter-intuitive, such situations are either harmless examples of nonstandard capacities or incoherent. This, in turn, means that proponents of different anthropic assumptions (like the Self-Indication Assumption of Dieks [1992] and Olum [2002], as well as many earlier authors), need to find other arguments for the suggested insufficiency of the Self-Sampling Assumption.[11]

Bibliography

Adams, Fred C. and Laughlin, Gregory 1997: 'A dying universe: the long-term fate and evolution of astrophysical objects' in *Reviews of Modern Physics*, 69, pp. 337-372.

Benford, Gregory, 1980: *Timescape*. New York: Pocket Books.

Bostrom, Nick, 1999: 'The Doomsday Argument is Alive and Kicking' in *Mind*, 108, pp. 539-550.

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

Bostrom, Nick, 2001: 'The Doomsday Argument Adam & Eve, UN+ +, and Quantum Joe' in *Synthese*, 127, pp. 359-387.

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

Brown, Bryson, 1992: 'Defending Backwards Causation' in *Canadian Journal of Philosophy*, 22, pp. 429-444.

Chalmers, David J., 1996: *The Conscious Mind*. Oxford: Oxford University Press.

Craig, William L., 1979, *The Kalam Cosmological Argument*. London: Macmillan.

Cramer, John G., 1983: 'The Arrow of Electromagnetic Time and the Generalized Absorber Theory' in *Foundations of Physics*, 13, 887-902.

Cirkovic. Milan M., 2003: 'Resource Letter PEs-1: Physical eschatology' in *American Journal of Physics*, 71, pp. 122-133.

Dieks, Dennis, 1992: 'Doomsday – Or: The Dangers of Statistics' in *Philosophical Quarterly*, 42, pp. 78-84.

Earman, John, 1995: *Bangs, Crunches, Whimpers, and Shrieks*. Oxford: Oxford University Press.

[11] The author is happy to acknowledge Nick Bostrom for many pleasant discussions, Miloš Arsenijevic and Petar Grujic for sparking his interest in the topic, Ministry of Science and Environment of Serbia for partial support, Ivan Havel for kindly allowing the use of his Figure 1, Jorge Luis Borges and Philip K. Dick for creative inspiration, and Irena Diklic for all her love and tender encouragement.

Flew, A. 1954, ‘Can A Cause Precede Its Effect?’ in *Proceedings of the Aristotelian Society Supplementary Volume* 28, pp. 45-62.

Gott, J. Richard, 1993: ‘Implications of the Copernican principle for our future prospects’ in *Nature,* 363, pp. 315-319.

Grünbaum, Adolf, 1973: *Philosophical Problems of Space and Time*. Dordrecht: Reidel.

Grünbaum, Adolf, 1991: ‘Creation as a Pseudo-Explanation in Current Physical Cosmology’ in *Erkenntnis,* 35, pp. 233-254.

Grünbaum, Adolf, 1998: ‘Theological Misinterpretations of Current Physical Cosmology’ in *Philo,* 1, Issue 1.

Grünbaum, Adolf, 2000, ‘A New Critique of Theological Interpretations of Physical Cosmology’ in *British Journal for the Philosophy of Science,* 51, pp. 1-43.

Havel, Ivan M., 1999: ‘Living in Conceivable Worlds’ in *Foundations of Science* **3**, 375-394.

Leslie, John, 1996: *The End of the World: The Ethics and Science of Human Extinction*. London: Routledge.

McArthur, Robert P. and Slattery, Michael P., 1974: ‘Peter Damian and Undoing the Past’ in *Philosophical Studies,* 25, pp. 137-141.

Mellor, David H., 1981: *Real Time* Cambridge: Cambridge University Press.

Price, Huw, 1996: *Time's Arrow and Archimedes' Point*. Oxford: Oxford University Press.

Olum, Ken, 2002: ‘The doomsday argument and the number of possible observers’ in *Philosophical Quarterly* 52, pp. 164-184.

Remnant, P. 1978, ‘Peter Damian: Could God Change the Past?’ in *Canadian Journal of Philosophy,* 8, pp. 259-268.

Smith, Quentin, 1988, ‘The Uncaused Beginning of the Universe’ in *Philosophy of Science,* 55, pp. 39-57.

Smith, Quentin, 1990: ‘A Natural Explanation of the Existence and Laws of Our Universe’ in *Australasian Journal of Philosophy,* 68, pp. 22-43.

Swinburne, R. 2000, ‘Discussion. Reply to Grünbaum’ in *British Journal for the Philosophy of Science,* 51, pp. 481-485.

Tappenden, Paul, 2000: ‘Identity and probability in Everett's multiverse’ in *British Journal for the Philosophy of Science,* 51, pp. 99-114.

Milan M. Cirkovic
Astronomical Observatory, Volgina 7,
11000 Belgrade, Serbia and Montenegro
e-mail: mcirkovic@aob.aob.bg.ac.yu