Refutations of the Simulation Argument

Fabien Besnard, Fabien.Besnard@wanadoo.fr

5th September 2004

Abstract

By examining the logical consistency of Nick Bostrom's simulation argument, we find that its conclusions are not fully compatible with its premises, a weak form of the liar's paradox. We also claim that a flaw in the simulation argument is to be found in the misuse of finite probability theory.

1 Introduction

The argument developed in [1], known as the simulation argument, has recently attracted much attention. Even some popular science magazines like [3] have accounted for it. According to this argument, a three-way disjonction must be true. More precisely, at least one of the following statements must be true : (1) almost no civilization of human level ever reach a stage of very advanced technology, called post-humanity and characterized by the ability to run a computer simulation of a whole human-level civilization, (2) almost all civilizations that reach post-humanity do not desire to run computer simulation of human-level civilizations, or (3) we probably live in a computer simulation.

Since the consequences of this argument are rather dramatic, it must be investigated in details. Authors have already criticized the argument on the ground on its rather wild implications, perhaps containing the seeds of its own invalidity. P. Davies even sees it as a *reductio ad absurdam* of multiverse theories ([4]). However, it must be noted that the conclusion of the simulation argument is not (3), but only a three-way disjunction involving (3). Is it enough to allow one to discard the argument in the name of rationality ? At any rate, since the simulation argument is claimed to be a rational argument, it is important to inspect its logical and mathematical aspects. We propose to tackle to this task in this article.

The simulation argument relies on several assumptions such as the substrate independence and the indifference principle that we do not wish to discuss and take for granted in the following. For convenience we summarize the simulation argument in the second section. Then, in the third section we use a strategy inspired by the liar's paradox to reduce the argument to the absurd. We conclude that there must be a flaw in the argument, and give two related propositions about it in the fourth section. We call them the cardinality problem and the viewpoint problem.

2 Summary of the simulation argument

The simulation argument consists in two parts. In the first, which we will call the foundation, the author, relying on several technical conjectures in computer science, claims that it is highly probable that an enormous quantity of computing power should be available to sufficiently advanced civilizations. These advanced civilizations are called 'post-human'. It is argued that post-human civilizations, or individuals in them, could use the immense computing power they possess to simulate human-level civilizations like ours, though only a fraction f_I of them would be interested in doing that. We call N_I the average number of simulations that each interested post-human civilization would then creates. Because of the enormous computing power of post-human civilization, N_I is huge. This is the conclusion of the foundation part of the argument.

In the second part, called 'the core', the author computes the fraction of all human-level civilizations that are being simulated. This is :

$$f_{sim} = \frac{f_P f_I N_I}{(f_P f_I N_I) + 1}$$

where f_P is the fraction of human-level civilization that reach post-humanity. Then, it is easily shown that at least one of the following is true :

$$f_p \approx 0 \qquad (1)$$

$$f_I \approx 0 \qquad (2)$$

$$f_{sim} \approx 1 \qquad (3)$$

The interpretation of (3) is that we probably live in a computer simulation.

The simulation argument can thus be summarized by the following diagram :

$$F \longrightarrow N_I >> 0 \Longrightarrow (1) \ or \ (2) \ or \ (3) \qquad (*)$$

The foundation is '$F \longrightarrow N_I >> 0$'[1], and the core is '$N_I \Longrightarrow (1)$ *or* (2) *or* (3)'. For our purpose, all we need to know is that F depends on the currently known laws of physics and some conjecture about the technical feasibility of very complex computer simulations.

3 Incompatibility of (3) with the foundation of the argument

Let us suppose that some piece of evidence (S) has been found that make us suspect that (1) and (2) are false, while (3) is true. Then let us look at the line of reasoning we get if we throw (S) in the hypotheses, living all other parts of the simulation argument unchanged :

$$F \text{ and } (S) \longrightarrow N_I >> 0 \Longrightarrow (1) \text{ or } (2) \text{ or } (3) \qquad (**)$$

Logically if some theory makes a prediction, and this prediction is realized, at the very least the confidence we have in the theory cannot lessen. But in fact (**) is a very doubtful argumentation. Indeed, given (S), the confidence we have in the laws of physics would be much lessened. In his article Bostrom emphasizes that the laws of physics may be very different than the ones we know of, if we happen to be simulated. Also, it is very possible that there is an automatized procedure that detects any human-made experience which would reveal an intrinsic granulosity of space or time ([1]). Actually such a granulosity would breaks Lorentz invariance, because a minimal length cannot be Lorentz-contracted. Recent theories of quantum gravity actually predict such a phenomenon, and upper bounds on a minimal unit of lenght have already been deduced from experiment ([2]). This means, according to Bostrom, that the results of these experiments could very well have been forged by 'those who simulate us'. It is easy to see that in such conditions, a radical form of scepticism should prevail. In particular, predictions about what could be achieved in the future would not be very reliable (if ever such predictions can be). Actually, even our belief that the sun will rise tomorrow would be diminished, since the simulation could be terminated at any moment. We claim that in view of this, the credence we can have in (**) must be necessarily less than the credence we can have in (*). Thus, (*) would turn out to be a theory that becomes less and less credible as its predictions are realized ! It does not mean that in

[1] We recall that $>> 0$ means 'very big'

such a situation we should not believe to be in a simulation, but that this belief would only be based on (S) and not at all on the simulation argument.

In view of the preceeding paragraph, it appears that the simulation argument cannot be correct if (1) and (2) are false. It means that among the eight possible triples of truth-values (1), (2) and (3) can take according to the conclusion of the simulation argument, at least one (false, false, true) is incompatible with the foundation part of this argument. This does not imply that the simulation argument is false, but at least it is incomplete. Let us look for example at the following deduction :

$$x^2 = 1 \text{ and } x > 0 \Longrightarrow x = 1 \text{ or } x = -1$$

It is perfectly true from a logical point of view, but obviously something has been forgotten because one part of the final disjonction is incompatible with one of the premise. We would say that this deduction is incomplete, and even misleading because it forces us to consider a case that cannot happen. Of course, once the incompatibility has been found, the deduction is easily refined by removing $x = -1$ from the conclusion.

By carefully inspecting the simulation argument, it appears that every step of the argumentation seems compatible with the hypothesis of (1) and (2) being false. Perhaps this is not so and we are not clever enough to see it. In this case, the simulation argument could be refined, thus leading to conclusion that (1) or (2) must be true. It means that the 'simulation part' of the simulation argument would have to be removed !

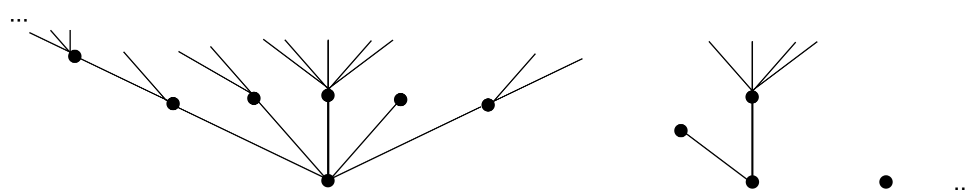
In any case, there must be a flaw somewhere in the simulation argument.

4 Possible flaws in the core of the argument

4.1 The cardinality problem

A fraction is the quotient of two numbers. To calculate f_{sim}, f_I and f_P, we need the cardinal of the set of all post-human civilizations, all human-level civilizations, all simulated human-level civilizations, all post-human civilizations interested in simulating human-level civilizations. The question that naturally arises is 'are these sets finite ?'. The biggest set that appears in the argument is the set of all civilizations whether human or post-human, simulated or not. In probability theory this set would be called the universe, but here it would be more appropriately called the multiverse. Let us call it Ω. It is not needed to know the precise structure of Ω as long as it is a finite set, because one just has to assume equiprobability of every element of Ω to do the calculation. On the other hand, if Ω might be infinite we would have to prescribe a certain probability law[2] on it, thus its structure would have to be known. We see that it is crucial for the argument that Ω be a finite set. In the following we will see that in some cases, it is not so.

At the end of his paper, N. Bostrom consider the possibility that a simulation could take place inside another simulation. Actually, this possibility must be considered, because it is a consequence of a recursive application of the simulation argument. Therefore the multiverse could have a 'forest-like' structure, like in the following figure.



A possible structure for the multiverse

The root of each tree represent a 'real' post-human civilization, and each node represents a simulated civilization, which can be post-human and simulate other civilizations as well. The root without branches represents a human non-simulated civilization. As N. Bostrom puts it, we cannot know for sure what the laws of physics at each node or root look like. If they are very different from the ones we know of, the possibility that a node may generate an infinity of branches cannot be ruled out, neither the possibility that there may be an infinite number of trees in the 'forest'. In either case Ω is infinite and the argument is flawed.

Suppose on the contrary that there are only a finite number of trees, let us say just one for the sake of simplicity. Then, for the tree to be finite, it is necessary that each node of the tree gives birth to a finite number of branches only and that there is no path of infinite length in the tree[3]. It could appear reassuring because if the root has only a finite amount of computing power, an assumption that sounds reasonable, it seems at first sight that it cannot generates an infinity of branches at the first level, neither

[2] or at least a measure.
[3] this is König's lemma of graph theory.

sustain an infinitely high tower of simulations. Now let us ask the question : 'what time is it ?'. It is a very important question because the definition of Ω does not refer to any notion of time. Thus, the root could run a simulation s_1, from its begining to its end, then a simulation s_2 and so on, and the resulting tree (representing all the simulations *ever* generated by the root) would have an infinity of branches. It would even be possible for the root to simulate an infinity of infinitely long lasting simulations, just by running s_1 for a unit of time, then s_1 and s_2 each for a unit of time, then s_1, s_2 and s_3 again for one time unit, and so on. The simulations would run more and more slowly but nothing could be suspected 'from within'. We do not find any reason to exclude the possibility that a post-human civilization live for ever. In this case, Ω is again infinite. In the end we find that having an infinite computing power or living for ever amounts roughly to the same for a post-human civilization.

While the above does not prove that Ω is infinite, it casts serious doubt on the possibility of doing the simple calculations of the 'core' of the simulation argument [4].

4.2 The viewpoint problem

There is an even worse problem in the definition of Ω : it is not obvious that it is a set ! To answer this question, we must wonder what is our point of view when we speak of all civilizations. This collection cannot of course be gathered before our eyes, even in principle. Therefore we cannot describe it as a part of the physical world. The only rational tool that remains is then pure mathematics. But we know since B. Russell that describing something by its properties is not a valid way of defining a set. For instance, the set of all sets does not exists, neither does the set of all sets that have some fixed mathematical structure (like the set of all vector spaces). The usual way of defining a set is to describe it as a subset of some known set. Since we lack a mathematical description of the multiverse, we cannot affirm that it is a set.

We think that the problem of viewpoint is the key to understand the flaw in the simulation argument. The latter begins with rather plausible assumptions about the world we know and love, which we follow without noticing a rather subtle change in the point of view, and at the end we have gotten completely off the ground. We think that for an argument aiming at calculating the probability that we live in some part P of a bigger set Ω,

[4]It should also be noted that not only Ω but several subsets of Ω appearing in the calculation may be infinite, and N_I as well in some cases.

the definition of Ω should be made clear from a mathematical point of view. This is a difficulty that cannot be avoided : it is well known in probability theory that the set of all possible outcomes must be prescribed before any calculation is undertaken. Once this problem has been raised, we see the rather wild implications of the mere existence of Ω and this, in our mind, forbids the argument to go any further. More simply put, if we imagine from the begining that we *may* be in a simulation, the argument cannot go further. At least it must be ammended.

5 Conclusion

The method we have used in section 3 makes the simulation argument appear very reminiscent of the liar's paradox, though in a weaker form that does not give rise to a full-fledged paradox. A statement like 'I can prove that nothing can be proved' is easily seen to be incompatible with its own conclusion and our method would apply to it. Like the simulation argument, it can be given the truth-value 'false'. However the simulation argument is not an argument of formal binary logic. As such, it need not have a truth-value. However it should be noted that in our method we have only made use of the reliability one could assign to the simulation argument, and shown that it would decrease if some of its predictions turned out to be realized.

One might be tempted to conclude likewise that any so-called rational argument whose conclusion would be incompatible with rational thinking is to be rejected. However, it should be noted that we have not claimed nor used such a strong principle in our refutation of the simulation argument. We have carefully split the argument in two parts : the foundation part that is based on physics and more generally on the natural sciences, and the core, which is based on mathematics. We have argued that the validity of the conclusion of the simulation argument is not compatible with its foundation part, but we have not claimed that it spoils any form of rational thinking, including mathematics, since it could be argued that it does not (a question that sounds like 'Does God have the power to make Pythagoras' theorem false ?').

It should also be noted that our criticism about the mathematical part of the simulation argument is logically independent from our rational questioning of section 3. Thus we believe to have provided two independent refutations of the simulation argument. However, it is very interesting to note that the inspection of the mathematical foundation of the argument reveal a 'viewpoint problem' that also looks like the 'liar's problem' of section

3.

Of course, the physical part of the argument by itself is certainly not immune to critcism. We believe that even if a simulation of the physical laws is conceivable to a certain extent, the complexity of the initial conditions for a good simulation is a problem that cannot be overestimated. However, we do not feel competent to argue on this point, and leave it to others.

6 References

References

[1] N. Bostrom, "Are you living in a computer simulation ?", *Philosophical Quarterly* **vol. 53** (2003), No 211, p. 243-255.

[2] G. Amelino-Camelina, "Are we at the dawn of quantum-gravity phenomenology ?", *Towards Quantum Gravity, Lecture Notes in Physics*, **541**, p.1-44

[3] J.P. Delahaye, "Sommes-nous réels ?", Pour la Science, August 2004.

[4] P. Davies, "Fitness and the cosmic environment", *in Fitness of the cosmos for life : Biochemistry and fine-tuning*, too appear.