The Simulation Argument and the Reference Class Problem : a Dialectical Contextualism Analysis

Paul FRANCESCHI
paul.franceschi@yahoo.fr

English postprint of a paper initially published in French in Philosophiques, 2016, 43-2, pp. 371-389, under the title *L'argument de la Simulation et le problème de la classe de référence : le point de vue du contextualisme dialectique*

ABSTRACT. I present in this paper an analysis of the Simulation Argument from a dialectical contextualist's standpoint. This analysis is grounded on the reference class problem. I begin with describing in detail Bostrom's Simulation Argument. I identify then the reference class within the Simulation Argument. I also point out a reference class problem, by applying the argument successively to three different reference classes: aware-simulations, imperfect simulations and immersion-simulations. Finally, I point out that there are three levels of conclusion within the Simulation Argument, depending on the chosen reference class, that yield each final conclusions of a fundamentally different nature.

1. The Simulation Argument

I shall propose in what follows an analysis of the *Simulation Argument*, recently described by Nick Bostrom (2003). I will first describe in detail the Simulation Argument (SA for short), focusing in particular on the resulting counter-intuitive consequence. I will then show how such a consequence can be avoided, based on the analysis of the reference class underlying SA, without having to give up one's pre-theoretical intuitions.

The general idea behind SA can be stated as follows. It is very likely that post-human civilizations will possess a computing power that will be completely out of proportion with that of ours today. Such extraordinary computing power should give them the ability to carry out completely realistic human simulations, such as ensuring that the inhabitants of these simulations are aware of their own existence, in all respects similar to ours. In such a context, it is likely that post-human civilizations will devote part of their computer resources to carrying out simulations of the human civilizations that preceded them. In this case, the number of simulated humans should greatly exceed the number of authentic humans. Under such conditions, taking into account the simple fact that we exist leads to the conclusion that it is more likely that we are part of the simulated humans, rather than of the authentic humans.

Bostrom thus points out that the Simulation Argument is based on the following three hypotheses:

(1) it is very likely that humanity will not reach a post-human stage
(2) it is very unlikely that post-human civilizations will carry out simulations of the human races that preceded them
(3) it is very likely that we are currently living in a simulation carried out by a post-human civilization

and it follows that at least one of these three assumptions is true.

For the purposes of the present analysis, it is also useful at this stage to emphasize the underlying dichotomous structure of SA. The first step in the reasoning consists then in considering, by dichotomy, that either (i) humanity will not reach a post-human stage, or (ii) it will actually reach such a post-human stage. The first of these two hypotheses corresponds to the disjunct (1) of the argument. We consider then the hypothesis that humanity will reach a post-human stage and thus continue its existence for many millennia. In such a case, it can also be considered likely that post-human civilizations will possess both the technology and the skills necessary to perform human simulations. A new dichotomy then arises: either (i) these post-human civilizations will not perform such simulations — this is the disjunct (2) of the argument; or (ii) these post-human civilizations will actually perform such simulations. In the latter case, it will follow that the number of simulated humans will greatly exceed the number of humans. The probability of living in a simulation will therefore be much greater than that of living in the shoes of an ordinary human. The conclusion then follows that we, the inhabitants of the Earth, are probably living in a simulation carried out by a post-human civilization. This last conclusion constitutes the disjunct(3) of the argument. An additional step leads then to the conclusion that at least one of the hypotheses (1), (2) and (3) is true. The dichotomous structure underlying SA can thus be described step by step as follows:

(4)	humanity will either not reach a post-human stage or reach a post-human stage	dichotomy 1
(1)	humanity will not reach a post-human stage	hypothesis 1.1
(5)	humanity will reach a post-human stage	hypothesis 1.2
(6)	post-human civilizations will be able to perform human simulations	from (5)
(7)	post-human civilizations will either not perform human simulations or will perform them	dichotomy 2
(2)	post-human civilizations will not perform human simulations	hypothesis 2.1
(8)	post-human civilizations will perform human simulations	hypothesis 2.2
(9)	the proportion of simulated humans will far exceed that of humans	from (8)
(3)	it is very likely that we are currently living in a simulation carried out by a post-human civilization	from (9)
(10)	at least one of the hypotheses (1), (2) and (3) is true	from (1), (2), (3)

It is also worth mentioning an element that results from the very interpretation of the argument. For as Bostrom (2005) points out, the Simulation Argument must not be misinterpreted. This is not an argument that leads to the conclusion that (3) is true, namely that we are currently living in a simulation carried out by a post-human civilization. The core of SA is thus that one of the hypotheses (1), (2) or (3) at least is true.

This nuance of interpretation being mentioned, the Simulation Argument is not without its problems. Because SA leads to the conclusion that at least one of the assumptions (1), (2) or (3) is true, and that in the situation of ignorance in which we find ourselves, we can consider the latter as equiprobable. As Bostrom himself notes (Bostrom, 2003): “In the dark forest of our current ignorance, it seems sensible to apportion one's credence roughly evenly between (1), (2) and (3)”. However, according to our pre-theoretical intuition, the probability of (3) is nil or at best extremely close to 0, so the conclusion of the argument has the consequence of increasing the probability that (3) is true from zero to a probability of about 1/3. Thus, the problem with the Simulation Argument is precisely that it shifts — via its disjunctive conclusion — from a zero or almost zero probability concerning (3) to a much higher probability of about 1/3. Because a probability of 1/3 for the hypotheses (1) and (2) is not a priori shocking, but is completely counter-intuitive as far as hypothesis (3) is concerned. It is in this sense that we can talk about the problem posed by the Simulation Argument and the need to find a *solution* to it.

As a preliminary point, it is worth considering what constitutes the paradoxical aspect of SA. What indeed gives SA a paradoxical nature? For SA differs from the class of paradoxes that lead to a contradiction. In paradoxes such as the Liar or the sorites paradox, the corresponding reasoning leads to a contradiction[1]. However, nothing of the sort can be seen at the level of SA, which belongs, from this point of view, to a different class of paradoxes, including the Doomsday Argument and Hempel's problem. It is indeed a class of paradoxes whose conclusion is contrary to intuition, and which comes into conflict with the set of all our beliefs. In the Doomsday Argument then, the conclusion that taking into account our rank within the class of humans who have ever existed has the effect that an apocalypse is much more likely than one might have initially thought, offends the set of all our beliefs. Similarly, in Hempel's problem, the fact that a blue umbrella confirms the hypothesis that all crows are black comes in conflict with the body of our knowledge. Similarly within SA, what finally appears paradoxical at first analysis is that SA leads to a probability of the hypothesis that we are currently living in a simulation created by post-humans, which is higher than that resulting from our pre-theoretical intuition.

2. The reference class problem and the Simulation Argument

The conclusion of the reasoning underlying SA, based on the calculation of the future ratio between real and simulated humans, albeit counter-intuitive, nevertheless results from a reasoning that appears a priori valid. However, such reasoning raises a question, which is related to the *reference class* that is inherent to the argument itself[2]. Indeed, it appears that SA has, indirectly, a particular reference class, which is that of human *simulations*. But what constitutes a simulation? The original argument implicitly refers to a reference class which is that of virtual simulations of humans, of a very high quality and by nature indistinguishable from authentic humans. However, there is some ambiguity about the very notion of a simulation and the question arises as to the applicability of SA to other types of human simulations[3]. Indeed, we are in a position to conceive of somewhat different types of simulations which also fall intuitively within the scope of the argument.

As a preliminary point, it is worth specifying here the nature of the simulations carried out by computer means referred to in the original argument. Implicitly, SA refers to computer simulations carried out by means of conventional computers composed of silicon chips. But it can also be envisaged that simulations are carried out using computers built from

[1] The Liar is thus both true and false. In the sorites paradox, an object with a certain number of grains of sand is both a heap and a non-heap. Similarly, in Goodman's paradox, an emerald is both green and grue, and therefore both green and blue after a certain date. Finally, in the Sleeping Beauty paradox, the probability that the piece fell on heads before the awakening of the Sleeping Beauty is 1/2 by virtue of one reasoning mode, and only 1/3 by virtue of an alternative reasoning.

[2] William Eckhardt (2013, p. 15) considers that — in the same way as the Doomsday Argument (Eckhardt 1993, 1997, Franceschi, 2009) — the problem inherent in SA comes from the use of retrocausality and the problem related to the definition of the reference class: "if simulated, are you random among human sims? hominid sims? conscious sims?".

[3] We will leave aside here the question of whether an infinite number of simulated humans should be taken into account. This could be the case if the ultimate level of reality were abstract. In this case, the reference class could include simulated humans who identify themselves, for example, with matrices of very large integers. But Bostrom answers such an objection in his FAQ (www.simulation-argument.com/faq.html) and points out that in this case, the calculations are no longer valid (the denominator is infinite) and the ratio is not defined. We will therefore leave this hypothesis aside, focusing our argument on what constitutes the core of SA, i.e. the case where the number of human simulations is finite.

components using DNA properties and molecular biology. Recent research has shown that it is possible to implement high-performance algorithms (Adleman 1994, 1998) and to produce computer components (Benenson & al. 2001, MacDonald & al. 2006) based on bio-calculation techniques that exploit in particular the combinations of the four components (adenine, cytosine, guanine, thymine) of the DNA molecule. If such a field of research were to expand significantly and make it possible to produce computers at least as powerful as conventional computers, this type of bio-computers could legitimately fall within the scope of SA as well. Because the fact that the simulations are carried out using conventional or biological computers[4] does not alter the scope of the argument. In any case, the result is that the proportion of simulated humans will be much higher than that of real humans, due to the properties of simulated reality using digital means, because the computer does not know the physical limits that are those of matter.

It can also be observed preliminarily that Bostrom explicitly refers to simulations carried out using computer means. However, the question arises as to whether simulated humans could not consist of perfectly successful physical copies of real humans. In such a case, simulations[5] could be extremely difficult to discern. A priori, such a variation also constitutes an acceptable version of SA. However, there is a difference with the original argument, which also highlights Bostrom's preferential choice of computer simulations. Indeed, in the original argument there is a very significant disproportion between humans simulated by computer means on the one hand and real humans on the other. This is the premise (9) of the argument: "the proportion of simulated humans will far exceed that of humans". As Bostrom points out, the former would then be much more numerous than the latter, due to the very nature of computer simulations. It is this disproportion that then allows us to conclude (3) "we most probably live in a simulation carried out by a post-human civilization". With simulations of a physical nature, one would not a priori have such a disproportion, and the scope of the conclusion would be somewhat different. Suppose, for example, that post-humans manage to perform simulations of a physical nature, the number of which would be equal to that of real humans. In this case, the proportion of simulated humans would be 1/2 (whereas it is close to 1 in the original argument). Premise (9) would then become: "the proportion of simulated humans and actual humans will be 1/2". And this would only allow us to conclude (3) "the probability that we are simulations performed by a post-human civilization is equal to 1/2". As can be seen, this would result in a significantly attenuated version of SA. The difference with the original version of SA is that the simulation argument for physical simulations applies with less force than the original argument. However, if the conditions were to change and this would result in the future in a disproportion of the same nature as with computer simulations for physical simulations, SA would then apply with all its force. In any event, the following analysis would then apply in the same way to this last category of simulations.

With these preliminary considerations in mind, we shall focus in turn on different types of human simulations, which are likely to be part of the SA reference class, and the ensuing conclusions at the argument level. Because the very question of defining the reference class for SA leads to questions about whether or not several types of simulations should be included within the SA scope. However, the question of the definition of the reference class for SA thus appears closely related to the nature of the future taxonomy of the beings and entities that will populate the Earth in the near or distant future. There is no question here of claiming exhaustiveness, given the speculative nature of such an area. However, it is possible to determine to what extent SA can also be applied to simulations of a different nature from those mentioned in the original argument, but which have equal legitimacy. We shall

[4] The same would be true if simulations were carried out using quantum computers.

[5] I thank an anonymous referee for highlighting this point, as well as the point about computers built from components using DNA properties and molecular biology.

examine then in turn: conscious simulations, imperfect simulations, and immersion simulations.

3. The reference class problem : the case of conscious simulations

At this step, it is not yet possible to really talk about the *problem* of the reference class within SA. To do so, it must be shown that the choice of one or the other reference class has completely different consequences at the level of the argument, and in particular that the nature of its conclusion is affected, i.e. fundamentally modified. In what follows, we will now focus on showing that depending on which reference class is chosen, radically different conclusions ensue at the level of the argument itself and that, consequently, there is a *reference class problem* within SA. For this purpose, we will consider several reference classes in turn, focusing on how conclusions of a fundamentally different nature result from them at the level of the argument itself.

The original version of SA implicitly depicts simulations of humans of a certain type. These are virtual simulations, almost indistinguishable from real humans and that present thus a very high degree of sophistication. Moreover, these are a type of simulations that are not aware that they are themselves simulated and are therefore convinced that they are genuine humans. This is implicit in the terms of the argument itself and in particular, the inference from (9) to (3) which leads to the conclusion that 'we' are currently living in an indistinguishable simulation carried out by post-humans. In fact, these are simulations that are somehow abused and misled by post-humans regarding their true identity. For the purposes of this discussion, we shall term *quasi-humans*$^-$ the simulated humans who are not aware that they are human.

At this stage, it appears that it is also possible to conceive of indistinguishable simulations that have an identical degree of sophistication but that, on the other hand, would be aware that they are being simulated. We shall then call *quasi-humans*$^+$ the simulated humans who are aware that they are themselves simulations. Such simulations are in all respects identical to the *quasi-humans*$^-$ to which SA implicitly refers, with the only difference that they are this time clearly aware of their intrinsic nature of simulation. Intuitively, SA also applies to this type of simulation. A priori, there is no justification for excluding such a type of simulation. Moreover, there are several reasons to believe that *quasi-humans*$^+$ may be more numerous than *quasi-humans*$^-$. For ethical reasons (i) first of all, it may be thought that post-humans might be inclined to prefer *quasi-humans*$^+$ to *quasi-humans*$^-$. For the fact of conferring an existence on quasi-humans constitutes a deception as to their true identity, whereas such an inconvenient is absent in the case of *quasi-humans*$^+$. Such deception could reasonably be considered unethical and lead to some form of prohibition of *quasi-humans*$^-$. Another reason (ii) is that simulations of humans who are aware of their own simulation nature should not be dismissed a priori. Indeed, we can think that the level of intelligence acquired by some quasi-humans in the near future could be extremely high and in this case, the simulations would very quickly become aware that they are themselves simulations. It may be thought that from a certain degree of intelligence, and in particular that which may be obtained by humanity in the not too distant future (Kurtzweil, 2000, 2005; Bostrom, 2006), quasi-humans should be able — at least much more easily than at present — to collect evidence that they are the subject of a simulation. Furthermore (iii), the very concept of "unconscious simulation that it is a simulation" could be inherently contradictory, because it would then be necessary to limit one's intelligence and therefore, it would no longer constitute an indistinguishable and sufficiently realistic simulation[6]. These three reasons suggest that *quasi-humans*$^+$ may well

[6] It seems difficult to rule out here the case where *quasi-humans*$^-$ discover, at least fortuitously, that they are simulated humans, thus becoming *quasi-humans*$^+$ from that moment on. However, in order to advantage the

exist in greater numbers than *quasi-humans*$^{-}$ — or even that they may even be the only type of simulation implemented by post-humans.

At this stage, it is worth considering the consequences of taking into account the *quasi-humans*$^{+}$ within the simulation reference class inherent to SA. For this purpose, let us first consider the variation of SA (let us term it SA*) that applies, exclusively, to the class of *quasi-humans*$^{+}$. Such a choice, first of all, has no consequence on the disjunct (1) of SA, which refers to a possible disappearance of our humanity before it has reached the post-human stage. Nor does this has any effect on the disjunct (2), according to which post-humans will not perform *quasi-humans*$^{+}$, i.e. conscious simulations of human beings. On the other hand, the choice of such a reference class has a direct consequence on the disjunct (3) of SA. Certainly, it follows, in the same way as for the original argument, the first level conclusion that the number of *quasi-humans*$^{+}$ will far exceed the number of authentic humans (the *disproportion*). However, the second level conclusion that "we" are currently quasi-humans no longer follows. Indeed, such a conclusion (let us call it *self-applicability*) no longer applies to us, since we are not aware that we are being simulated and are completely convinced that we are authentic humans. Thus, in this particular context, *the inference from (9) to (3) no longer prevails*. Indeed, what constitutes SA's *worrying* conclusion no longer results from step (9), since we cannot identify with the *quasi-humans*$^{+}$, the latter being clearly aware that they are evolving in a simulation. Thus, unlike the original version of SA based on the reference class that associates humans with *quasi-humains*$^{-}$, this new version associating humans with *quasi-humans*$^{+}$ is not associated with such a disturbing conclusion. The conclusion that now follows, as we can see, is quite *reassuring*, and in any case very different from the deeply *worrying*[7] conclusion that results from the original argument.

At this stage, it appears that a question arises: should we identify, in the context of SA, the reference class to the *quasi-humans*$^{-}$ or the *quasi-humans*$^{+}$?[8] It appears that no objective element in SA's statement supports the a priori choice of the *quasi-humans*$^{-}$ or the *quasi-humans*$^{+}$. Thus, any version of the argument that includes the preferential choice of the *quasi-humans*$^{-}$ or the *quasi-humans*$^{+}$ appears to be biased. This is the case for the original version of SA, which thus contains a bias in favor of the *quasi-humans*$^{-}$, which results from Bostrom's choice of a class of simulations that is exclusively assimilated to *quasi-humans*$^{-}$, i.e. simulations that are not aware of their simulation nature and are therefore abused and misled by post-humans about the very nature of their identity. And this is also the case for SA*, the alternative version of SA that has just been described, which includes a particular bias in favor of *quasi-humans*$^{+}$, simulations that are aware of their own simulation nature. However, the choice of the reference class is fundamental here, because it has an essential consequence: if we choose a reference class that associates simulations with quasi-humans, the result is the *worrying* conclusion that we are most likely currently experiencing in a

paradox, we will consider here that the very notion of an indistinguishable simulation is not plagued with contradiction.

[7] Bostrom (2003) considers that the fact that we live in a simulation would only moderately affect our daily lives: "Supposing we live in a simulation, what are the implications for us humans? The foregoing remarks notwithstanding, the implications are not all that radical". However, it may be thought that the effect should be much more profound, given that the fundamental level of reality is not where the simulation subjects believe it to be and that, as a result, many of their beliefs are completely erroneous. As David Chalmers (2005) points it out: "The brain is massively deluded, it seems. It has all sorts of false beliefs about the world. It believes that it has a body, but it has no body. It believes that it is walking outside in the sunlight, but in fact it is inside a dark lab. It believes it is one place, when in fact it may be somewhere quite different".

[8] For the purposes of this discussion, we present things as an alternative between *quasi-humans*$^{-}$ and *quasi-humans*$^{+}$. However, one could conceive that post-humans – perhaps different post-human civilizations – create both *quasi-humans*$^{-}$ and *quasi-humans*$^{+}$. We would then have a tripartite situation involving humans, *quasi-humans*$^{-}$ and *quasi-humans*$^{+}$. For the sake of simplicity, we can assimilate here such a situation to the one that prevails when post-humans only create *quasi-humans*$^{-}$ since it is sufficient that the latter are present in very large numbers to create the worrying effect inherent to SA.

simulation. On the other hand, if a reference class is chosen that identifies simulations with *quasi-humans*$^+$, the result is a scenario that *reassuringly* does not include such a conclusion. At this stage, it is clear that the choice of the *quasi-humans*$^-$ i.e., non-conscious simulations — in the original version of SA, to the detriment of conscious simulations, constitutes an arbitrary choice. Indeed, what makes it possible to prefer the choice of *quasi-humans*$^-$, compared to *quasi-humans*$^+$? Such justification is lacking in the context of the argument. At this stage, it appears that SA's original argument contains a bias that leads to the preferential choice of *quasi-humans*$^-$, and to the alarming conclusion associated with it[9].

4. The reference class problem : the case of imperfect simulations

The problem of the reference class within SA relates, as mentioned above, to the very nature and to the type of simulations referred to in the argument. Is this problem limited to the preferential choice, at the level of the original argument, of unconscious simulations, to the detriment of the alternative choice of conscious simulations, which correspond to very sophisticated simulations of humans, capable of creating illusion, but endowed with the awareness that they themselves are simulations? It appears not. Indeed, as mentioned above, other types of simulations can also be envisaged for which the argument also works, but which are of a somewhat different nature. In particular, it is conceivable that post-humans may design and implement simulations that are identical to those of the original argument, but that are not as perfect in essence. Such a situation is quite likely and does not have the ethical disadvantages that could accompany the indistinguishable simulations staged in the original argument. The choice to carry out such simulations could be the result of the necessary technological level, or of deliberate and pragmatic choices, designed to save time and resources. These could be, for example, simulations of excellent quality such that the scientific inhabitants of the simulations could only discover their artificial nature after, for example, ten years of research. Such simulations could be carried out in very large numbers and, given their less resource-intensive nature, could occur in even greater numbers than *quasi-humans*$^-$. For the purposes of this discussion, we will call *imperfect* simulations this category of simulations.

At this stage, one can ask oneself what are the consequences on SA of taking into account a reference class that identifies itself with *imperfect simulations*? In this case, it follows, in the same way as the original argument, that the first level conclusion that the number of *imperfect simulations* will far exceed the number of authentic humans (the *disproportion*). But here too, however, the second level conclusion that “we” are currently *imperfect simulations* (*self-applicability*) no longer follows. The latter no longer applies to us and a reassuring conclusion replaces it, since we are clearly aware that we are not such *imperfect simulations*. Finally, it turns out that the conclusion that results from taking into account the class of imperfect simulations is of the same nature as that which follows when considering the class of the *quasi-humans*$^+$.

5. The reference class problem : the case of immersion simulations

As we have seen, extending the SA reference class to conscious simulations leads to a conclusion of a different nature from the one that results from the original argument. The same applies to another category of simulations — imperfect simulations — which lead to a

[9] This type of bias can be analyzed in one instance of the *one-sidedness bias* (Walton, 1999, p. 76-81, Franceschi, 2014, p. 587-592) where the reference class is that of the simulations and the associated duality is consciousness/unconsciousness.

conclusion of the same nature as conscious simulations, and which in any case turns out to be different from that resulting from taking into account the simulations mentioned in the original argument. At this stage, the question arises as to whether the reference class can not be assimilated to other types of simulations relevant from the point of view of SA and whose consideration would lead to a conclusion that is inherently different from that which follows when considering the simulations of the original argument, or conscious or imperfect simulations.

In particular, the question arises as to whether human simulations, which would be such as to apply to ourselves — in a sense that may differ from the original argument — and which would include the conclusion of *self-applicability* inherent in SA, could not exist in a more or less near future. Some answers can be provided by considering an evolution of the concepts of virtual reality that are already being implemented in different fields such as psychiatry, surgery, industry, military training, entertainment, etc. In psychiatry in particular, virtual universes are used to implement techniques related to behavioral therapies, and offer advantages over traditional in vivo scenarios (Powers & Emmelkamp, 2008). In this type of treatment, the patient himself is simulated using an avatar and the universe in which he evolves is also simulated in the most realistic way possible. Convincing results have been obtained in the treatment of some phobias (Choy & al., 2007, Parsons & Rizzo, 2008), as well as post-traumatic stress disorder (Cukor & al., 2009, Baños & al., 2011).

In this context, it is conceivable that developments in this concept of virtual reality could lead to the realization of simulated humans, which would require a high degree of realism. This would require, in particular, the completion of current research, particularly on the simulation of the human brain. It is possible that significant progress may be made in the near future (Moravec, 1998; Kurzweil, 2005; Sandberg and Bostrom, 2008; De Garis et al. al., 2010). It is also conceivable that we will then have the ability to immerse ourselves in simulated universes by borrowing the personalities of humans thus simulated, while really having — the time of immersion — the impression that this is our real existence[10]. In addition, the same human simulation could take the form of multiple variations that would correspond to the purpose — therapeutic, scientific, playful, utilitarian, historical, etc. — sought during the immersion. For example, it is conceivable that some variations may only include important elements of the simulated personality's life, neglecting uninteresting details. For the purposes of this discussion, we can term this type of simulation: *immersion simulations*. In this context, humans could thus frequently resort to immersion in a simulated anterior human personality. It is also possible that individuals may use simulations of themselves: they could be simulations of themselves at earlier times in their lives, with eventual slight variations, however, depending on the purpose sought for the immersion in question. In such circumstances, it is conceivable that very large quantities of this type of simulation could be carried out by computer means. In any case, it appears that the number of simulations at our disposal would be much greater than the inhabitants of our planet. In this context, it appears that SA functions in the same way as the original argument if we reason in relation to a reference class that identifies itself with this type of *immersion simulations*.

At this point, it is worth considering the effect on SA of assimilating the reference class to *immersion simulations*. In such a context, it appears that the first-level consequence based on the humans/simulations *disproportion* would apply here, in the same way as the original argument. Secondly, and this is an important consequence, the second level conclusion based on *self-applicability* would now apply, since we can conclude that "we" are also, in this extended sense, simulations. On the other hand, it would no longer follow the alarming

[10] A complete simulation of a human brain is also called an *upload*. One definition (Sandberg & Bostrom, 2008, p. 7) is as follows: : "The basic idea is to take a particular brain, scan its structure in detail, and construct a software model of it that is so faithful to the original that, when run on appropriate hardware, it will behave in essentially the same way as the original brain."

conclusion, which is that of the original argument and which manifests itself at a third level, that we are unconscious simulations, since the fact that we are in this sense simulations does not imply here that we are mistaken about our first identity. Thus, unlike the original argument, the result is a *reassuring* conclusion: humans are occasionally *immersion simulations*, while being aware that they use them.

Could we not object here that we have not yet reached the state where we can identify, even if only temporarily, with such *immersion simulations* and that this does not make the above developments relevant to SA? Strictly speaking, the virtual reality implemented in our time can indeed be considered too coarse in nature to be assimilated to the very realistic simulations hinted at by Bostrom. However, it can be assumed that only high-quality *immersion simulations*, which would give the illusion at least the time of their use that they are a real existence, could be carried out, for such simulations to become relevant for the SA reference class. The hypothesis that such a technological level, based on an explosion of artificial intelligence, could be achieved within a few decades has thus been put forward (Kurzweil, 2005; Eden et al. al., 2013). If such a technological evolution were to occur within, for example, a few decades, could we not then legitimately consider that such simulations also fall within the reference class of SA? Given this possible temporal proximity, it seems appropriate to take into account the case of *immersion simulations* and to evaluate their consequences for SA[11].

6. The different levels of conclusion according to the chosen reference class

Finally, the preceding discussion emphasizes that if SA is considered in light of its inherent reference class problem, there are actually several levels in the conclusion of SA: (C1) disproportion; (C2) self-applicability; (C3) unconsciousness (the worrying fact that we are fooled, deceived about our primary identity). In fact, the previous discussion shows that (C1) is true regardless of the chosen (by restriction or extension) reference class: quasi-humans$^-$, quasi-humans$^+$, imperfect simulations and immersion simulations. In addition, (C2) is also true for the original reference class of quasi-humans$^-$ — and for immersion simulations, but is false for the class of quasi-humans$^+$ and imperfect simulations. Finally, (C3) is true for the original reference class of quasi-humans$^-$, but it proves to be false for quasi-humans$^+$, imperfect simulations and immersion simulations. These three levels of conclusion are represented in the table below:

[11] The above also shows that when examining SA carefully, it can be seen that the argument contains a second reference class. This second reference class is that of *post-humans*. What is a post-human? Should we assimilate this class to civilizations far superior to ours, to those that will evolve in the 25th century or the 43th century? Should the descendants of our current human race who will live in the 22nd century be counted among the post-humans if they were to make considerable technological progress in the field of simulations? In any case, the definition of the post-human class appears to be closely linked to that of simulations. Because if we are interested, in a broad sense, in *immersion simulations*, then post-humans can be assimilated to a generation of humans not very far from us. If we consider *imperfect simulations*, then they should be associated with a more distant time. On the other hand, if we consider, in a more restrictive sense, simulations of humans that are completely indistinguishable from our current humanity, then we should be interested in post-humans from a much more distant era. Thus, the class of post-humans appears to be closely correlated with that of simulations, because the degree of evolution of simulations is related to the level reached by the post-human civilizations that implement them. For this reason, we shall limit the present discussion to the reference class of the simulations.

level	*conclusion*	*case*	*quasi-humans*$^-$	*quasi-humans*$^+$	*imperfect simulations*	*immersion simulations*
C1	the proportion of simulated humans will far exceed that of humans (*disproportion*)	C1A	true	true	true	true
	the proportion of simulated humans will not significantly exceed that of humans	C1Ā	false	false	false	false
C2	we are most likely simulations (*self-applicability*)	C2A	true	false	false	true
	we are most likely not simulations	C2Ā	false	true	true	false
C3	we are unconscious simulations of their simulation nature (*unconsciousness*)	C3A	true	false	false	false
	we are not unconscious simulations of their simulation nature	C3Ā	false	true	true	true

Figure 1. The *different levels of conclusion within SA*

as well as in the following tree structure:

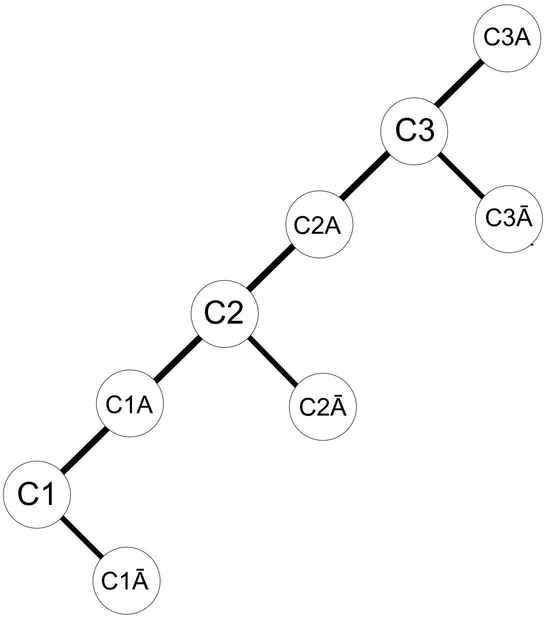

Figure 2. *Tree of the different levels of conclusion of SA*

While SA's original conclusion suggests that there is only one level of conclusion, it turns out, however, as just pointed out, that there are in fact several levels of conclusion in SA, when the argument is examined from a broader perspective, in the light of the reference class problem. The conclusion of the original argument (C3A) is itself worrying and alarming, in that it concludes that there is a much higher probability than we had imagined a priori that we are humans simulated without being aware of it. However, the above analysis shows that, depending on the chosen reference class, some conclusions of a very different nature can be inferred by the simulation argument. Thus, a completely different conclusion is associated with the choice of the reference class of the *quasi-humans*$^+$ or *imperfect simulations*. The resulting conclusion is that we are not such simulations (C2Ā). Finally, another possible conclusion, itself associated with the choice of the *immersion simulation* class, is that we are eventually part of such a simulation class, but we are aware of it and therefore it is not a cause for concern (C3Ā).

The above analysis finally highlights what is wrong with the original version of SA, which is at a twofold level. First, the original argument focuses on the class of simulations that are not aware of their own simulation nature. This leads to a succession of conclusions that there will be a greater proportion of simulated humans than authentic humans (C1A), that we are part of simulated humans (C2A) and finally that we are, more likely than we might have imagined a priori, simulated humans unaware of being (C3A). However, as mentioned above, the very notion of human simulation is ambiguous, and such a class can in fact be defined in different ways, given that there is no objective criterion in SA for choosing such a class in a way that is not arbitrary. We can indeed choose the reference class by identifying the simulations with *unconscious* simulations, i.e. *quasi-humans*$^-$ simulations. But the alternative choice of a reference class that identifies itself with simulations that are *conscious* of being simulations themselves, i.e. *quasi-humans*$^+$, has equal legitimacy. In the original argument, there is no objective criterion for choosing the reference class in a non-arbitrary way. Thus, the fact of favoring, in the original argument, the choice of *quasi-humans*$^-$ — with the alarming conclusion associated with them — over *quasi-humans*$^+$, constitutes a *bias*, as well as the choice of a reference class that identifies itself with *quasi-humans*$^+$, leads this time to a reassuring conclusion.

Secondly, it appears that the reference class of SA can be defined at a certain level of restriction or extension. The choice in the original argument of the *quasi-humans*$^-$ — occurs at a certain level of restriction. But if we now move to a certain level of extension, the reference class now includes *imperfect simulations*. And if we place ourselves at an even greater level of extension, simulations include not only imperfect simulations, but also *immersion simulations*. But depending on whether the class is chosen at a particular level of restriction or extension, a completely different conclusion will follow. Thus, the choice, at a higher level of extension, of imperfect simulations leads to a reassuring conclusion. Similarly, at an even greater level of extension, which this time includes immersion simulations, there also follows a new reassuring conclusion. Thus, the above analysis shows that in the original version of SA, the choice is made preferentially, by restriction, on the reference class of *quasi-humans*$^-$, to which is associated a worrying conclusion, as well as a choice by extension, also taking into account *imperfect simulations* or *immersion simulations*, leads to a reassuring conclusion.

Can we not object, at this stage, that the above analysis leads to a change in the original scenario of SA and that it is no longer the same problem[12]? To this, it can be replied that the previous analysis is based on variations in SA that preserve the very structure of the original argument. What this analysis shows is that this same structure is likely to produce conclusions of a very different nature, as long as the reference class is varied within reasonable limits that correspond to the context of SA, and even though the original SA statement suggests a single type of conclusion. Bostrom himself emphasizes that it is the structure of the argument that constitutes its real core: “The *structure* of the Simulation

[12] I thank an anonymous referee for raising this objection.

Argument does not depend on the nature of the hypothetical beings that would be created by the technologically mature civilizations. If instead of computer simulations they created enormous numbers of brains in vats connected to a suitable virtual reality simulation, the same effect could in principle be achieved." (Bostrom, 2005). In addition, the different levels of extension used here to highlight variations in the SA reference class are intended to illustrate how different levels of conclusion can result. But if we wish to preserve the very form of the original argument, we can then limit the variation of the reference class to what really constitutes the core of this analysis, by considering only a reference class that identifies itself with the quasi-humans. The reference class is then made up of both quasi-humans$^-$ and quasi-humans$^+$. This is sufficient to generate a reassuring conclusion — which is not taken into account in the original argument — and thus modify the general conclusion resulting from the argument. In this case, it is the same reference class as the one underlying the original argument, with the only difference that simulations knowing that they are simulated are now part of it. Because the latter, whose possible existence is not mentioned in the original argument, nevertheless have an equal right to legitimacy in the context of SA.

Finally, the preferential choice in the original argument of the *quasi-humans*$^-$ class, appears to be an arbitrary choice that no objective criterion justifies, while other choices deserve equal legitimacy. For the SA statement does not contain any objective element allowing the choice of the reference class to be made in a non-arbitrary manner. In this context, the worrying conclusion associated with the original argument also turns out to be an arbitrary conclusion, since there are several other reference classes that have an equal degree of relevance to the argument itself, and from which a quite reassuring conclusion follows.[13] [14]

References

Adleman Leonard « Molecular Computation of Solutions to Combinatorial Problems », Science, vol. 266, 1994, p. 1021-1024.

Adleman Leonard « Computing with DNA », Scientific American, vol. 279(2), 1998, p. 54-61.

Baños R.M., Guillen V. Quero S., García-Palacios A., Alcaniz M., Botella C. «A virtual reality system for the treatment of stress-related disorders», International Journal of Human-Computer Studies, vol. 69, no. 9, 2011, p. 602–613.

Benenson Y., Paz-Elizur T., Adar R., Keinan E., Livneh Z., Shapiro E. «Programmable and autonomous computing machine made of biomolecules», Nature, vol. 414, 2001, p. 430–434.

Bostrom, Nick « Are You a Living in a Computer Simulation? », Philosophical Quarterly, vol. 53, 2003, p. 243-55.

Bostrom, Nick « Reply to Weatherson », Philosophical Quarterly, vol. 55, 2005, p. 90-97.

[13] The resulting double weakening of SA finally makes it possible to reconcile SA with our pre-theoretical intuitions, because the worrying scenario of the original argument now coexists with several scenarios of a quite reassuring nature.

[14] The present analysis is a direct application to the Simulation Argument of the form of *dialectical contextualism* described in Franceschi (2014).

I thank two anonymous referees for *Philosophiques*, for very useful comments on an earlier version of this article.

Bostrom, Nick « How long before superintelligence? », Linguistic and Philosophical Investigations, vol. 5, no. 1, 2006, p. 11-30.

Chalmers, David « The Matrix as Metaphysics », dans Grau C., dir., Philosophers Explore the Matrix, New York, Oxford University Press, 2005.

Choy Yujuan, Fyer A., Lipsitz J., « Treatment of specific phobia in adults », Clinical Psychology Review, vol. 27, no. 3, 2007, p. 266–286.

Cukor Judith, Spitalnick J., Difede J., Rizzo A., Rothbaum B. O., « Emerging treatments for PTSD », Clinical Psychology Review, vol. 29, no. 8, 2009, p. 715–726.

Franceschi, Paul, « A Third Route to the Doomsday Argument », Journal of Philosophical Research, vol. 34, 2009, p. 263-278.

Franceschi, Paul (2014), « Eléments d'un contextualisme dialectique », in J. Dutant, D. Fassio & A. Meylan, dir., Liber Amicorum Pascal Engel, Genève, Université de Genève, p. 581-608.

De Garis, Hugo, Shuo, C., Goertzel, B., Ruiting, L., « A world survey of artificial brain projects, part i: Large-scale brain simulations », Neurocomputing, vol. 74, no. 1-3, 2010, p. 3-29.

Eckhardt, William, « Probability Theory and the Doomsday Argument », Mind, vol. 102, 1993, p. 483-488.

Eckhardt, William, « A Shooting-Room View of Doomsday », Journal of Philosophy, vol. 94, 1997, p. 244-259.

Eckhardt, William, Paradoxes in probability Theory. Dordrecht, New York, Springer, 2013.

Eden A., Moor J., Søraker J., Steinhart E. (eds.) Singularity Hypotheses: A Scientific and Philosophical Assessment, Londres, Springer, 2013.

Kurzweil, Ray, The Age of Spiritual Machines: When Computers Exceed Human Intelligence, New York & Londres, Penguin Books, 2000.

Kurzweil, Ray, The Singularity is Near, New York, Viking Press, 2005.

MacDonald J., Li Y., Sutovic M., Lederman H., Pendri K., Lu W., Andrews B. L., Stefanovic D., Stojanovic M. N. « Medium Scale Integration of Molecular Logic Gates in an Automaton », Nano Letters, 6, 2006, p. 2598–2603.

Moravec, Hans « When will computer hardware match the human brain? », Journal of Evolution and Technology, 1998, vol. 1.

Parsons T.D., Rizzo A. « Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis », Journal of Behavior Therapy and Experimental Psychiatry, vol. 39, no. 3, 2008, p. 250–261.

Powers M. B., Emmelkamp P. « Virtual reality exposure therapy for anxiety disorders: A meta-analysis », Journal of Anxiety Disorders, vol. 22, no. 3, 2008, p. 561–569.

Sandberg, Anders et Bostrom, Nick Whole Brain Emulation: a Roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University, 2008.

Walton, Douglas, One-Sided Arguments: A Dialectical Analysis of Bias, Albany, State University of New York Press, 1999.