The Simulation Argument and the Reference Class Problem: the dialectical contextualist's standpoint

Paul Franceschi
University of Corsica
p.franceschi@univ-corse.fr

Paul Franceschi
Fontaine du salario
Lieu-dit Morone
20000 Ajaccio
France

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 Bostrom's Simulation Argument step-by-step. I identify then the reference class within the Simulation argument. I also point out a reference class problem, by applying the argument successively to several references classes: aware-simulations, rough-simulations and cyborg-type 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 a solution to solve the problem posed by the *Simulation argument*, recently described by Nick Bostrom (2003). I shall first attempt to describe in detail the Simulation argument, by exposing in particular its inherent problem. I will show then how a solution can be brought to such a problem, based on the analysis of the reference class which underlies the Simulation argument, and without it being necessary to give up one's pretheoretical intuitions.

The general idea which underlies the Simulation argument (SA, for short) can be expressed as follows. It is very likely that post-human civilizations will possess a calculus computing power completely out of proportion with the one which is ours at present time. Such an extraordinary computing power should confer them the capacity to implement completely realistic human simulations, such in particular that the inhabitants of these simulations would be conscious of their own existence, in every respect similar to ours. In such a context, we can think that it is likely that post-human civilizations will actually dedicate a part of their computing resources to realize simulations of the human civilizations which preceded them. In this case, the number of the simulated human beings should very largely encompass that of the genuine human beings. In such conditions, the fact of taking into account the mere fact that we exist leads to the conclusion that it is more likely that we belong to the simulated human beings, rather than to the genuine ones.

Bostrom also manages to describe the Simulation argument accurately. He underlines that SA is based on the following three hypotheses:

(1)	humanity will face a nearest extinction
(2)	the post-human civilizations will not realize human beings' simulations
(3)	we currently live in a simulation realized by a post-human civilization

The first step of the reasoning consists in considering, by dichotomy, that either (i) humankind will face a nearest extinction, or (ii) she will pursue its existence in the distant future. The first of these two hypotheses constitutes the disjunct (1) of the argument. We consider then the hypothesis according to which the humanity will not face a nearest extinction and will then pursue its existence through numerous millenniums. In such a case, we can also consider that it is likely that the post-human civilizations will possess at the same time the technology and the required capacities to realize human beings' simulations. A new dichotomy then follows: either (i) these post-human civilizations will not realize such simulations—it consists of the disjunct (2) of the argument; or (ii) these post-human civilizations will actually realize such simulations. In this last case, it will follow that the number of simulated human beings will largely exceed that of the human beings. The probability to live in a simulation will thus be much greater than that to live as an ordinary human being. It follows then the conclusion that we, inhabitants of the Earth, probably live in a simulation realized by a post-human civilization. This last conclusion constitutes the disjunct (3) of the argument. An additional step leads then to consider that in the lack of any evidence in favour of the one or the other of them, we can consider the hypotheses (1), (2) and (3) as equiprobable.

The Simulation argument can thus be described step-by-step as follows:

(4)	either humankind will face a nearest extinction, or humankind w not face a nearest extinction	dichotomy 1
(1)	humankind will face a nearest extinction	hypothesis 1.1
(5)	humankind will not face a nearest extinction	hypothesis 1.2
(6)	the post-human civilizations will be capable of realizin human beings' simulations	from (5)
(7)	either the post-human civilizations will not realize huma beings' simulations, or they will realize them	dichotomy 2
(2)	the post-human civilizations will not realize human being simulations	hypothesis 2.1
(8)	the post-human civilizations will realize human being simulations	hypothesis 2.2
(9)	the proportion of the simulated human beings will ve largely exceed that of the human beings	from (8)
(3)	we currently live in a simulation realized by post-human civilization	from (9)
(10)	in the lack of evidence in favour of one of them, the hypotheses ((2) and (3) are equiprobable	from (1), (2), (3

It is also worth mentioning an element which results from the interpretation of the argument. For as Bostrom himself (2005) notes, the Simulation argument must not be wrongly interpreted. It is not indeed an argument which leads to the conclusion that (3) is true, namely that we currently live in a simulation realized by a post-human civilization. The core of the Simulation argument lies then in the fact that the propositions (1), (2) or (3) are equiprobable.

This nuance of interpretation being mentioned, the Simulation argument does not however miss to raise a *problem*. For the argument leads to the conclusion that one of the propositions (1), (2) or (3) at least is true, and that in the situation of ignorance where we are, we can consider them as equiprobable. As Bostrom himself puts it: “In the dark forest of our current ignorance, it seems sensible to apportion one's credence roughly evenly between (1), (2) and (3).” (Bostrom 2003). However, according to our pre-theoretical intuition, the probability of (3) is null or at best extremely close to 0. So, the conclusion of the argument has for consequence to make pass the probability that (3) is the true, from zero to a probability of about 1/3. So, the problem posed by SA is precisely that it makes shift—via its disjunctive conclusion—a probability of zero or of near zero concerning (3) to a much more considerable probability of about 1/3. For a probability of 1/3 concerning proposition (1) and (2) has nothing shocking a priori, but reveals itself on the other hand completely counter-intuitive as regards proposition (3). It is in this sense that we can speak of the problem posed by the Simulation argument and of the need of looking for a *solution* to the latter.

In a preliminary way, it is worth wondering about what constitutes the paradoxical nature of SA. What is it indeed that confers a paradoxical aspect to SA? For SA distinguishes itself from the class of paradoxes that lead to a contradiction. In the paradoxes such as the Liar or the sorites paradox, the corresponding reasoning leads to a contradiction: the Liar is both true and false. In the sorites paradox, an object containing a certain number of grains of sand is both a heap and a non-heap. No such thing manifests itself at the level of SA which belongs, from this point of view, to a different class of paradoxes of which is also part the Doomsday argument. It consists indeed of a class of paradoxes the conclusion of which presents a counter-intuitive nature, and comes in conflict with the whole set of our beliefs. In the Doomsday argument, the conclusion according to which the fact of taking into consideration our own birth rank within the class of the human beings having never existed results in the fact that a doomsday is much more likely than we would have possibly envisaged initially, comes to strike the set of all our beliefs. In a similar way, what appears finally here as paradoxical, in first analysis, is that SA leads to a probability of the hypothesis according to which we currently live in a simulation created by post-humans, which is greater than the one which results from our pre-theoretical intuition.

2. The reference class within the Simulation Argument

The conclusion of the reasoning which underlies SA, based on the calculation of the future ratio between the real human beings and the simulated ones, as it proves to be counter-intuitive, results nevertheless from a reasoning which seems a priori valid. However, such a reasoning leads to an interrogation, which is associated with the *reference class* which is inherent to the argument itself.[1] Indeed, it turns out that SA contains, in an indirect way, a specific class of reference, which is that of the human beings' *simulations*. But what is it then that constitutes a simulation? The original argument refers, in an implicit way, to a reference class which is that of virtual simulations of human beings, of a very high quality and by nature indiscernible from the genuine ones. However, a certain ambiguity lies in the mere

[1] William Eckhardt (2013, p. 15) considers that—in the same way as with the Doomsday argument (Eckhardt 1993, 1997, Franceschi 2009)—the problem inherent to SA results from the use of reverse-causality and from the problem associated with the definition of the reference class: 'if simulated, are you random among human sims? hominid sims? conscious sims?'.

notion of simulation and the question arises of the applicability of SA to other types of human beings' simulations.[2] We can indeed conceive of somewhat different types of simulations which, in an intuitive way, also enter the scope of the argument.

It is possible to imagine, first, a type of simulations in every respect identical to those described in the original argument, i.e. almost indiscernible from genuine human beings, but with the only difference that they would be aware of their own nature of simulation. The only difference with the type of simulation described in the original argument would thus be that these last simulations would clearly be conscious of being not authentic human beings. A priori, nothing excludes that the post-humans would choose to implement any such simulations and intuitively, SA is also susceptible of applying to this particular type of simulations.

In the same way, SA refers implicitly to sophisticated simulations, of very high quality, which are by nature indiscernible from authentic human beings. However, we can conceive of various degrees in the quality of the human simulations. So the question notably arises of whether we can include in the reference class of SA some virtual simulations of a very slightly lower quality? With any such simulations, the nature of simulation which constitutes their deep identity would be susceptible of being one day discovered by the very subject. If the argument has to apply to this class of simulations, the question then arises of its applicability to other types of simulations of this nature, because we can conceive of numerous intermediate degrees between on the one hand, the indiscernible simulations and on the other hand, the simulations which we are currently capable of realizing, notably by means of computer generated images. So, the question does arise of whether the reference class of SA can go as far as to include simulations of lower quality than those evoked in the original argument?

Finally, it appears that SA also ‘works’ if we apply it to human beings whose brain is interfaced with *uploads*, simulations of the human mind including memorized events, knowledge, personality's traits, ways of reasoning, etc. relative to a given individual. We can imagine indeed that in a not very distant future, the emulation of the human brain could be achieved (Moravec 1998, Sandberg & Bostrom 2008, Garis & al. 2010), so that the realization of *uploads* could become common and intensively used. A very large number of *uploads* could be so realized and used in different purposes: scientific, cultural, social, utilitarian, etc. If we assimilate then the *uploads* to the simulations of SA, the argument also works. In a sense, the human beings endowed with *uploads* can be considered as simulations of partial nature, which only concern the brain or a part of the brain, even though the rest of the human body remains authentic and not simulated. In such a case, the human beings of which only the brain is simulated by means of an *upload*, can be assimilated to a particular type of *cyborgs*. We can so raise the general question of to what extent the class of the simulations of SA can be widened to partial simulations and to the types of cyborgs who have just been described. We can indeed conceive of cyborgs of various types, depending on the parts of the body and organs of replacement or substitution which are theirs. So the question does arise of to what extent SA also applies to this type of cyborgs?

As we can see it, the very question of the definition of the reference class for SA leads to wonder about the inclusion or not within the scope of SA of several types of simulations.

[2] We shall set aside here the question of whether or not we need to take into account an infinite number of simulated human beings. Such could be the case if the ultimate level of reality was abstract. In this case, the reference class could include simulated human beings who would identify themselves, for example, with matrices of very large integer numbers. But Bostrom answers such an objection in his FAQ (www.simulation-argument.com/faq.html) and indicates that in this case, the calculations do not apply any more (the denominator is infinite) and the ratio is not defined. We shall thus leave aside this hypothesis, by concentrating our argumentation on what constitutes the core of SA, i.e. the case where the number of human beings' simulations is finite.

Without pretending to be exhaustive, we can mention at this stage, among the latter: aware-simulations, the more or less rough-simulations and the partial simulations of a cyborg-type. The question of the definition of the reference class for SA seems then closely related to the nature of the future taxonomy of the beings and creatures which will populate the Earth in a near or distant future.

At this step, it turns out in first approach that the types of human beings' simulations present a somewhat varied nature, and that we can define the reference class of the simulations in several ways. We could then choose the reference class in a more or less restrictive or extensive way. In this context, it is worth delving more deeply into the consequences of the one or the other choice.

3. The reference class problem: the aware-simulations case

At this stage, we still cannot speak veritably of a reference class *problem* within SA. For it indeed, we need to show that the choice of the 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 finds itself modified in a fundamental way. In what follows, we shall from now on attempt to show that according to the choice of the one or the other reference class, some radically different conclusions follow at the level of the very argument and that consequently, there exists well a *reference class problem* within SA. For this purpose, we shall consider successively several reference classes, by attaching ourselves to show that some conclusions of a fundamentally different nature result from them at the level of the argument itself.

The original version of SA stages implicitly human beings' simulations of a certain type. They consist of virtual type simulations, almost indiscernible for ourselves and which present then a very high degree of sophistication. More still, they consist of a type of simulations which are not aware that they are themselves simulated and which are thus persuaded to be genuine human beings. This results implicitly of the terms of the argument itself and in particular, of the inference from (9) to (3) which leads to conclude that 'we' currently live in an indiscernible simulation realized by the post-humans. In fact, it consists of simulations which are somewhat abused and deceived by the post-humans as regards their true identity. For the needs of the present discussion, we shall term *quasi-humans*$^-$ the simulated human beings who are not aware that they are themselves simulated.

At this step, it turns out that we can also conceive of indiscernible simulations which present a completely identical degree of sophistication but which, on the contrary, would be aware that they are simulated. We shall then term *quasi-humans*$^+$ those simulated human beings who are aware that they are themselves simulations. Such simulations are in every respect identical to the *quasi-humans*$^-$ to which SA refers implicitly, with the only difference that they are this time clearly aware of their intrinsic nature of simulation. In an intuitive way, SA also applies to this type of simulation. A priori, we lack the justification to exclude such a type of simulations. More still, several reasons lead to think that the *quasi-humans*$^+$ could be more numerous than the *quasi-humans*$^-$. For ethical reasons (i) first, we can think that the post-humans could be inclined to prefer *quasi-humans*$^+$ to *quasi-humans*$^-$. For the fact of conferring an existence to the *quasi-humans*$^-$ constitutes a deceit on their real identity, while such an inconvenient is absent with regard to *quasi-humans*$^+$. Such a deceit could be reasonably considered as unethical and lead to one or the other form of *quasi-humans*$^-$ interdiction. Another reason (ii) militates for the fact of not pushing aside a priori those human beings' simulations that are aware of their own nature of simulation. We can think indeed that the level of intelligence acquired by certain *quasi-human* beings in a near future could be extremely high and make that in this case, the simulations would become very quickly aware that they are themselves simulations. We can think that starting from a certain degree of intelligence, and in particular the one susceptible of being obtained by humankind in a not very distant future (Kurtzweil 2000, 2005, Bostrom 2006), the *quasi-humans* should be able—at least much more easily than at present—to collect the evidence that they are the object of a simulation. Moreover, the concept of 'simulation which is unaware that it is a simulation' could be plagued with contradiction, because it would then be necessary to limit its intelligence and from then on, it would not consist any more of an indiscernible and enough realistic simulation. These two reasons suggest that the *quasi-humans*$^+$ could well exist in greater number than the *quasi-humans*$^-$.

At this stage, it turns out to be necessary to envisage the consequences of the consideration of the *quasi-humans*$^+$ within the reference class of the simulations inherent to SA. For that purpose, let us consider first the variation of SA (let us call it SA*) which applies, in an exclusive way, to the class of the *quasi-humans*$^+$. Such a choice has no consequence, first, on the disjunct (1) of SA, which refers to a possible next disappearance of our humanity. It has no effect either on the disjunct (2), according to which the post-humans will not realize *quasi-humans*$^+$, i.e. aware-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 with the original argument, the first-level conclusion according to which the number of *quasi-humans*$^+$ will largely exceed the number of genuine human beings (the *disproportion*). However, from now, the second-level conclusion according to which 'we' currently are *quasi-humans*$^+$, does not follow any more. Indeed, such a conclusion (let us term it the *self-applicability*) does no longer apply to us from now on, since we are not conscious of being simulated and are fully convinced of being genuine human beings. In effect, what constitutes the *disturbing* conclusion of SA does not result any more from now on from step (9), for we cannot identify ourselves with the *quasi-humans*$^+$, the latter being clearly aware that they live in a simulation. So, unlike the original version of SA based on the reference class which associates the human beings with the *quasi-humans*$^-$, this new version associating the human beings and the *quasi-humans*$^+$, is not related to such a disturbing conclusion. The conclusion which follows from now on, as we can see it, turns out to be completely *reassuring*, and in any case very different from that, profoundly *disturbing*, which results from the original argument.

At this step, it turns out that a question arises: must we identify, in the context of SA, the reference class with the *quasi-humans*$^-$ or with the *quasi-humans*$^+$? It turns out that no objective element, in the statement of SA, comes to justify the a priori choice of the *quasi-humans*$^-$ or of the *quasi-humans*$^+$. So, any version of the argument which contains the preferential choice of either the *quasi-humans*$^-$ or the *quasi-humans*$^+$ can be considered as exemplifying a *bias*. Such is then the case for the original version of SA, which contains then a bias in favour of the *quasi-humans*$^-$, which results from the choice by Bostrom of a class of simulations which assimilates itself exclusively with the *quasi-humans*$^-$, i.e. simulations which are unaware of their nature of simulations and which are consequently abused and deceived by the post-humans on the true nature of their identity. And such is also the case for SA* the alternate version of SA which has been just described, which contains a specific bias in favour of the *quasi-humans*$^+$, i.e. simulations that are aware of their own nature of simulation. However, the choice of the reference class proves here to be fundamental, for it contains an essential consequence: if we choose a reference class which associates the human beings with the *quasi-humans*$^-$, it results from it the *disturbing* conclusion that we currently very probably live in a simulation. On the other hand, if we choose a reference class which associates the human beings with the *quasi-humans*$^+$, if follows a scenario which in a *reassuring* way, does not entail such a conclusion. At this stage, it appears that the choice of the *quasi-humans*$^-$, i.e. unaware-simulations, in the original version of SA, to the detriment of aware-simulations, constitutes an arbitrary choice. In effect, what allows to prefer the choice of the *quasi-humans*$^-$ over the *quasi-humans*$^+$? Such a justification is lacking in the context of the argument. At this stage, it turns out that the original argument of SA contains a bias which leads to the preferential choice of the *quasi-humans*$^-$, and to the alarming conclusion which is associated with it. This remark being made, it is worth considering now the problem under a still wider perspective, by taking into consideration other possible types of simulations.

4. The reference class problem: the rough-simulations case

The reference class problem within SA bears, as mentioned above, on the very nature and the type of simulations implemented within the argument. Does this problem limit itself to the preferential choice, at the level of the original argument, of unaware-simulations, to the detriment of the alternate choice of aware-simulations, which correspond to human beings' very sophisticated simulations, capable of creating the illusion, but endowed with the awareness that they are themselves simulations? It seems not. Indeed, as mentioned above, we can also conceive of other types of simulations for which the argument also works, but which are themselves of a slightly different nature. In particular, we can imagine that the post-humans will conceive of and implement simulations that are identical to those of the original argument, but who do not however present a so perfect character. Such a situation presents a completely likely nature and does not present the ethical drawbacks which could accompany the indiscernible simulations hinted at in the original argument. The choice of realizing this type of simulations could result either from the required technological level, or from deliberate and pragmatic choices, intended to save time and resources. We can so conceive of various degrees in the realization of such type of simulations. It could involve for example simulations of very good quality, the artificial nature of which our current scientists could only discover after, say, ten years of research. But in an alternative way, such simulations could be of average quality, even rather rough, with regard to the above-mentioned almost indiscernible simulations. For the sake of the present discussion, however, we shall call *rough-simulations* this whole category of simulations.

What are then the consequences on SA of the consideration of a reference class which assimilates itself with rough-simulations? In such circumstances, a large number of such simulations would be detectable by us human beings. In this case, the first-level consequence based on the human beings/simulations *disproportion* always applies, in the same way as with the original argument. On the other hand, the second level conclusion based on the *self-applicability* does not apply any more now. For we can no longer conclude from now on that ‘we’ are simulations, since in the presence of any such simulations, we would quickly notice that they consist of simulated human beings and are not genuine human beings. So, in such a case, it turns out that the alarming conclusion inherent to the original version of SA and based on the *self-applicability* does no longer apply. A reassuring conclusion substitutes indeed itself to it, based on the fact that we human beings do not belong to this type of simulations.

At this stage, it appears that SA, in its original version, opts for the preferential choice of very sophisticated, undetectable simulations by us human beings and unaware of their nature of simulation. But as mentioned above, we can conceive of other types of simulations, of a more unrefined nature, to which the argument also applies. Until what level of detectable simulation can we go? Do we have to go as far as including in the reference class, at a higher level of extension, rather unrefined simulations, such as for example some improved versions of the simulations that which we are already capable of realizing by means of computer generated images? In this case, this leads to a slightly different formulation from the original argument, for we can then assimilate the class of the post-humans to the human beings who will live on Earth in ten years, or even in one year, or even—at a much greater level of extension—in one month. In this case, the disjunct (1) according to which humankind will not last until this time does no longer prevail, since such a technological level has already been reached. Also, the disjunct (2) has no longer any raison d'être, since we already realize such unrefined simulations. Thus, there only remains in this case the disjunct (3), which constitutes then the unique proposition which underlies the argument and constitutes the first-level conclusion of SA, according to which the number of the simulated human beings will largely encompass that of the genuine human beings. In this case, it follows well, in an identical way as with the original argument, the first-level conclusion according to which the number of *quasi-humans*$^+$ will largely exceed the number of genuine human beings (the *disproportion*). But there also, the second-level conclusion according to which 'we' currently are *quasi-humans*$^+$ (the *self-applicability*) does no longer follow. The latter does no longer apply to us from now on and a conclusion of reassuring nature substitutes itself to it, since we are clearly aware of being not such rough-simulations.

5. The reference class problem: the cyborg-case

As evoked above, another question that arises is whether the reference class can be widened to the cyborgs and in particular to this category of cyborgs who are indiscernible from human beings. We can indeed conceive of various types of cyborgs, going from those for which some parts of the body have been replaced by synthetic organs of substitution or more powerful, to those for which almost all organs—including the brain—have been replaced. A priori, such a class also enters the field of the argument. Here, the argument applies naturally to those elaborate, indiscernible from human beings cyborgs, for which a large part of the original organs have been replaced or transformed. In particular, the cyborgs for which a part of the brain was replaced by a—*partial* or not—*upload* enters naturally the field of the argument. Partial *uploads* are the ones for which only one part of the brain has been replaced by an *upload*. We can also imagine numerous types of *uploads* of this kind: *uploads* which reconstitute the memory by restoring the forgotten events can be so envisaged. They can prove themselves useful not only for healthy people, but also for those who suffer from diseases in which the memory functions are altered. We can conceive of that such types of partial *uploads* could be implemented in a more or less close future (Moravec 1998, Kurzweil 2005, Garis & al. 2010). And in the same way as with the original argument, we can conceive of that very large quantities of these *uploads* could be realized by computing means. In a general way, it turns out that the discussion about the inclusion of the cyborgs within the reference class of SA has its importance, because if we consider the class of the cyborgs in a wide sense, we are nearly already all cyborgs. If we indeed consider that organs or parts of the human body that have been replaced or improved so that they work correctly makes us cyborgs, such is today already the case, given the generalization of synthetic teeth, pacemakers, prostheses, etc. So the question arises of up to which degree we can include certain types of cyborgs within the scope of the argument.

What would then be the effect on SA of taking into account the class of the partial cyborgs, if we place ourselves at such a degree of extension? As well as for the rough-simulations, it turns out that the disjunct (1) according to which the human beings will not reach until this time does no longer prevail, since such a technological stage is already levelled off. In an identical way, the disjunct (2) does not justify itself any more either, because in such a context, we are already almost all any such partial cyborgs. So, the disjunct (3) only remains in this case as a unique proposition, but which emerges however under a different form from that of the original argument. In effect, the first-level consequence based on the human beings/simulations *disproportion* also applies here, in the same way as with the original argument. In addition, and it is an important difference here, the second-level conclusion based on the *self-applicability* also applies, since we can conclude from it that 'we' are also, in this wide sense, simulations. On the other hand, the alarming conclusion of the original argument that we are unaware-simulations, which manifests itself at a third level, does no longer follow, since the fact that we are simulations in this sense does not involve here that we are deceived on our true identity. So there finally ensues, unlike the original argument, a *reassuring* conclusion: we are simulations, who are fully aware of their own nature of partial cyborgs.

What precedes also shows that by examining SA with attention, we can notice that the argument holds a *second reference class*. This second reference class is that of the *post-humans*. What is then a post-human being? Must we assimilate this class to those civilizations that are very widely superior to ours, to those who will evolve either in the XXVth century or in the XLIIIth century? Must the descendants of our current human race who will live in the XXIIth century be counted among the post-humans? The fact that important evolutions associated with the increase of human intelligence (Moravec 1998, Kurzweil 2005) can arise in a more or less close future, constitutes in particular an argument which supports this assertion. But must we go as far as including the descendants of the current human beings who will live on Earth in 5 years? Such questions arise and require an answer. The question of how we have to define the post-humans, also constitutes then an element of the reference class problem of SA. In any case, the definition of the *post-humans*' class seems closely related to that of the *simulations*. For if one considers, in a wide sense, those cyborgs hardly more evolved than we currently are in a certain sense, then the post-humans can be assimilated with the human beings' next generation. The same goes if we consider rough-simulations slightly improved with regard to those that we are currently capable of producing. On the other hand, if we consider, in a more restrictive sense, simulations that are completely indiscernible from our current humanity, it is then worth considering post-humans of a clearly more distant time. In any case, it appears here that the reference class of the *post-humans*, as well as the class of the *simulations* which is associated with it, can be chosen at different levels of restriction or of extension.

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

Finally, the foregoing discussion emphasizes the fact that if we consider SA in the light of its inherent reference class problem, there are in reality several levels in the conclusion of SA: (C1) the disproportion; (C2) the self-applicability; (C3) the unawareness (the worrying fact that we are abused, deceived on our true identity). In fact, the previous discussion shows that (C1) is true whatever the chosen reference class (by restriction or by extension): the *quasi-humans*$^-$, the *quasi-humans*$^+$, rough-simulations and cyborg-type simulations. In addition, (C2) is also true for the original reference class of the *quasi-humans*$^-$ and for that of cyborg-type simulations, but proves however to be false for the class of the *quasi-humans*$^+$ and also for that of rough-simulations. Finally, (C3) is true for the original reference class of the *quasi-humans*$^-$, but proves to be false for the *quasi-humans*$^+$, rough-simulations and cyborg-type simulations. These three levels of conclusion are represented on the table below:

level	*conclusion*	*case*	*quasi-humans^{-}*	*quasi-humans^{+}*	*rough simulations*	*cyborg-type simulations*
C1	the proportion of the simulated human beings will largely exceed that of the human beings (*disproportion*)	C1A	true	true	true	true
	the proportion of the simulated human beings will not largely exceed that of the human beings	C1Ā	false	false	false	false
C2	very probably, we are simulations (*self-applicability*)	C2A	true	false	false	true
	very probably, we are not simulations	C2Ā	false	true	true	false
C3	we are simulations which are unaware of their nature of simulation (*unawareness*)	C3A	true	false	false	false
	we are not simulations which are unaware of their nature of simulation	C3Ā	false	true	true	true

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

and on the following tree structure:

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

Even though the original conclusion of SA suggests that there is only one single level of conclusion, it turns out however, as it has been revealed, that there are in reality several levels of conclusion within SA, inasmuch as we examine the argument from a wider perspective, in the light of the reference class problem. The conclusion of the original argument is itself disturbing and alarming, in the sense that it concludes to a much stronger probability that we had imagined it a priori, that we are simulated human beings who are not aware of it. Such a conclusion results from the path C1-C1A-C2-C2A-C3-C3A of the above tree. However, the preceding analysis shows that according to the chosen reference class, some conclusions of a very different nature can be inferred from the simulation argument. Hence, a conclusion of a completely different nature is associated with the choice of the reference class of the *quasi-humans*$^+$, but also of that of *rough-simulations*. The resulting conclusion is that we are not such simulations (C2Ā). This last conclusion is associated with the path C1-C1A-C2-C2Ā of the above tree. Finally, another possible conclusion, itself associated with the choice of the class of cyborg-type simulations, is that we belong to such a class of simulations, but that we are aware of it and that it presents thus nothing disturbing (C3Ā). This last conclusion is represented by the path C1-C1A-C2-C2A-C3-C3Ā.

Finally, the foregoing analysis casts light on the flaw in the original version of SA. The original argument focuses indeed on the class of simulations which are not conscious of their own nature of simulation. It follows then a succession of conclusions according to which there will be a larger proportion of simulated human beings than genuine human beings (C1A), that we belong to the simulated human beings (C2A) and finally that we are, more probably than we would have a priori imagined it, simulated human beings unaware of their being simulated (C3A). However, as evoked above, the very notion of human beings' simulation—itself associated with the class of the post-humans—shows itself ambiguous, and such a class can be defined in reality by different ways, given that there does not exist, within SA, an objective criterion allowing to choose such a class non-arbitrarily. In effect, we can choose the reference class by *restriction*, by identifying the simulations with the *quasi-humans*$^-$, or with the *quasi-humans*$^+$; in this case, the post-humans are the ones to which refers the original argument, of a much more distant time than ours. On the other hand, if we place ourselves at certain level of extension, the simulations assimilate themselves with less perfect simulations than those referred to in the original argument, as well as those of cyborg-type containing evolved *uploads*; in such a case, the associated post-humans are the ones of a less distant time. Finally, if we make the choice of the class of reference at a greater level of extension, the simulations are rough-simulations, hardly better than we are at present capable of realizing, or cyborg-type simulations with a degree of integration of simulated parts slightly much-evolved than the one that we know at present; in such a case, the class of associated post-humans is that of the human beings who will succeed us by a few years. As we can see it, we can make the choice of the reference class underlying SA at different levels of restriction or of extension. But according to whether the class will be chosen at such or such level of restriction or extension, a completely different conclusion will follow. So, the choice by restriction of perfect simulations that are unaware of their nature of simulation, as the argument original makes it, leads to a disturbing conclusion. On the other hand, the choice at slightly greater level of extension, of perfect simulations but aware of their nature of simulation, leads to a reassuring conclusion. And also, the choice, at a still greater level of extension, of rough-simulations or of cyborg-type simulations, also entails a reassuring conclusion. Hence, the preceding analysis shows that in the original version of SA, the choice concerns in a preferential way, by restriction, the reference class of the *quasi-humans*$^-$, with which a worrying conclusion is associated, whereas a choice by extension, taking into account the *quasi-humans*$^+$, *rough-simulations*, *cyborg-type* simulations, etc., leads to a reassuring conclusion. Finally, the preferential choice in the original argument of the class of the *quasi-humans*$^-$, appears then as an *arbitrary* choice that nothing comes to justify, whereas there exist several other classes that deserve an equal legitimacy. For there is no objective element in the statement of SA allowing to make the choice of the reference class non-arbitrarily. In this context, the disturbing conclusion associated with the original argument also turns out to be an arbitrary conclusion, whereas there exists several other reference classes which possess an equal degree of relevance with regard to the argument itself, and from which ensue a completely reassuring conclusion.[3]

References

Bostrom, N. (2003) Are You a Living in a Computer Simulation?, *Philosophical Quarterly*, 53, 243-55

Bostrom, N. (2005) Reply to Weatherson, *Philosophical Quarterly*, 55, 90-97

Bostrom, N. (2006) 'How long before superintelligence?', *Linguistic and Philosophical Investigations*, 5-1, 11-30

De Garis, H.D., Shuo, C., Goertzel, B., Ruiting, L. (2010) A world survey of artificial brain projects, part i: Large-scale brain simulations, *Neurocomputing*, 74(1-3), 3-29

Eckhardt, W. (1993) 'Probability Theory and the Doomsday Argument', *Mind*, 102, 483-88

Eckhardt, W. (1997) 'A Shooting-Room View of Doomsday', *Journal of Philosophy*, 94, 244-259

Eckhardt, W. (2013) *Paradoxes in probability Theory*, Dordrecht, New York : Springer

Franceschi, P. (2009) A Third Route to the Doomsday Argument, *Journal of Philosophical Research*, 34, 263-278

Franceschi, P. (2014) Eléments d'un contextualisme dialectique, in *Liber Amicorum Pascal Engel*, edited by J. Dutant, D. Fassio & A. Meylan, 581-608, English translation under the title *Elements of Dialectical Contextualism*, cogprints.org/9225

Kurzweil, R. (2000) *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, New York & London: Penguin Books

Kurzweil, R. (2005) *The Singularity is Near*, New York : Viking Press

Moravec, H. (1998) When will computer hardware match the human brain?, *Journal of Evolution and Technology*, vol. 1

Sandberg, A & Bostrom, N. (2008) Whole Brain Emulation: a Roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University

[3] The present analysis constitutes a direct application to the Simulation argument of the form of *dialectical contextualism* described in Franceschi (2014).