Consciousness, artificial intelligence. Transhumanism: truths and false promises

Michel Troublé - Ph.D - research director, robotics and artificial intelligence.

The current dominant thinking about the nature of living things and the cognition with which they are equipped is that their functionalities must all be reduced to « algorithms ». That is to say, sets of operating rules, instructions, applying to the development of more or less complex physico-chemical interactions such as oxygen and hydrogen gases that combine to form water.

We would thus be highly sophisticated machines, which would have spontaneously self-constructed by having the physically paradoxical capacity, given the entropic degradation of the universe, to ensure the durability of their structure in all circumstances, what characterizes them. Over time, these living machines would have developed a more or less developed intelligence through *self-learning* information processing systems.

The *consciousness* that we all naturally possess and which has the essential virtue of opening us to the colourful and sensitive perception of a universe otherwise closed on itself in its minerality, would then be only be an epiphenomenon, an accessory term that has no place in *transhumanist* thought that is only interested in the living in mechanical or computational terms. Essentially: repair, improve, these so-called living structures as we build rockets and computers are built, with ever-increasing technical performance.

The proponents of *transhumanist* thinking speculate that these various technical processes will one day be perfectly mastered, we should then be able to free ourselves from the physical and mental limitations acquired during our evolution. And that is why human beings could thus become much more intelligent and even almost immortal.

The ultimate goal of this mutation to the *transhuman* is to eliminate all physical and mental pain mechanisms, the downloading of the mind to a supercomputer could thus be the final solution to all our torments. To do this, it would be enough to copy on a 'key' of great capacity the different states of the neurons of our brain and then transfer them to the memory of a super-computer.

While it is not questionable that the tremendous technical progress in biology and neuroscience suggest that most diseases will disappear and that all or almost all parts of the body can be repaired or replaced, this purely computational approach to life that would lead to an infinite increase in our mental capacities is, however, not scientifically based.

It is the functional analysis of 'robots' that would be fully *autonomous,* i.e. artificially *alive* because able to ensure in all circumstances the durability of their structure, that will provide us with the formal elements that will justify this assertion.

A robot is a more or less complex physico-chemical system combining mechanical, electronic and computer elements. It is equipped with the following elements : *sensors* that connect the robot to the external environment, *actuators* that ensure its mobility and finally a *controller*, a calculation structure that establishes appropriate informational links between these sensors and actuators.

Autonomous robotics

The *autonomy* of robots is currently a major issue in terms of automation. The point is no less than to create physicochemical structures which, in the long term, would be alive artificially and as such endowed with capabilities similar to those of living beings in terms of the decisions they must take to ensure their durability in all circumstances. which characterizes them.

This *autonomy* is a recurrent demand for space exploration, given that the transmission time of a radio communication between the Earth and a vehicle that moves for example on the ground of Mars, can be as long as 20 minutes. Which thus prohibits any instantaneous control with the vehicle which must then get by on his own. It would therefore be desirable for this exploration robot to be totally *autonomous* as a human being could be in such circumstances.

Depending on the nature of the different objects that an *autonomous* robot would perceive with its sensors, the actions it would perform should therefore be such that they should ensure the durability of its structure, avoiding in particular to damage the delicate and very expensive measuring devices with which it is equipped.

If the robot is, for example, sensitive to any temperature increase of its structure, it would be necessary that regardless of the temperatures T measured by the thermometric sensor with which it is equipped, which temperatures would correspond to the hot objects encountered during its movements, its 'control' module or robot brain would be able to *create* two distinct *categories of actions* :

'Escape' or 'continue to move in the same direction' so that the temperature of its structure for example always less than about 30°C. This would then correspond to the creation of the following two *coherent categories of actions* : {'flee' if T>30°} but {'move forward' if T<30°}.

It should be noted that this functional analysis of a robot that would be *autonomous* applies equally to any physico-chemical structure such as that of a technician observing the movements of the mercury column of a thermometer with which it is equipped would decide to move carefully away from all the very hot objects he might encounter.

The indistinguishability of the world objects

This essential ability of the *autonomous* robot would therefore have to 'escape' a hot object if the temperature of its structure were to exceed 30°C would then logically imply that the different temperatures measured by the robot thermal sensor during its various displacements would obviously be *distinguishable* from each other by its controller.

Otherwise, the various robot movings initiated by the controller could only be performed *at random*, which would certainly be contrary to the expected ability to avoid any destructive approach with hot objects in its environment.

That the different temperatures displayed by thermometer sensor of the robot are physically *distinguishable* by its controller so that the robot can systematically turn away from all the very hot objects in its environment seems to go naturally.

We naturally perform a similar operation when in the morning we look at the temperature displayed by the thermometer hanging outside our house in order to know whether or not we should keep a warm clothing to get out of the house.

But this task which is very natural for us, is in fact paradoxically *infeasible* for any inanimate physical structure such as the robot.

We can indeed demonstrate mathematically from the formal theory of « pattern recognition »[1], which concerns the identification of object shapes based on their characteristic properties, that the different temperatures measured by the thermometric sensor of the robot are in this case *indistinguishable* by its controller which commands the locomotion system. That's what, for short, we will call the 'indistinguishability theorem'. Without being a rigorous demonstration of this theorem, the following scenario allows us to get very close to it :

Let's suppose that a gardener has two objects, a large stone that is heavy and a piece of wood that is light.

□ At first, the gardener wants to drive a nail that exceeds a board could hurt him. To do this, he uses the stone that is *heavy* and not the piece of wood that is obviously too *light* - it is his past experience that dictates this behaviour. The stone is thus a *different* object from the piece of wood in terms of the *action* of driving a nail.

□ In a second time, the same gardener decides to locate several spots in his garden to arrange plants. He then uses either the stone or the piece of wood to identify distinct locations in the garden. The stone is now an object *similar* to the piece of wood since they can have both the same marking function.

What must be remembered from this story is that the physical actions that can be associated with the two objects, utilitarially called 'stone' and 'piece of wood', essentially depend on the « gardener's good will ». In the absence of the gardener who is an already *living* physico-chemical structure, the two objects are fundamentally *indistinguishable* with regard to the different *actions* they can lead to. This is what the 'indistinguishability theorem' implies.

This primordial state of *indiscernibility* of material entities has until now been completely ignored by researchers for whom the *distinguishability* of macroscopic or microscopic objects

[1] Satosi Watanabe - *Pattern recognition, human and mechanical*, John Wiley & Son, 1985.

perceived/measured was self-evident and that it was therefore not necessary to question the validity of such an assertion.

This 'indistinguishability theorem' therefore calls totally into question the representation we make of the world and the way we act on it to ensure our durability, in other words to be *alive*.

To illustrate the considerable consequences implied by this 'indistinguishability theorem' with regard to *actions* we take in the world to ensure the durability of our existence, let us take the example of a simple thermostat whose sensor is a thermometer where the mercury meniscus is, for simplicity, either in the 'high' position (we then say that it is 'hot' in the room) or in the 'low' position (we then say that it is 'cold' in the room). The position of the meniscus being observed by an electrical device, a relay for example, that controls the start or stop of a heating device.

What the 'indistinguishability theorem' tells us in this case is that in the absence of a technician, which is a living physico-chemical structure, which would have *prepared* the thermostat by connecting together the various components in an appropriate way, the 'high' and 'low' positions of the mercury meniscus would be physically *indistinguishable* by its actuator, here an electrical relay. This would mean that the closing and opening of the relay that controls room heating could only be done at random. And not as desired by the person who buys the thermostat for his personal use : « it is 'cold', the heating system must be switched on », or « it is 'hot', the heating system must be switched off ».

This example of the thermostat is like the martian exploration robot which, in the absence of a technician, could only direct itself or turn away *at random* from all the objects it would encounter. Reactions that would ineluctably lead him to its destruction.

Darwinian natural selection

It seems that there is nevertheless a pragmatic way to make this exploration robot *autonomous*, it is to be inspired by the mechanism implemented in « darwinian natural selection » which, according to biologists would precisely explain the emergence of animation functions with which certain physico-chemical structures are equipped and then qualified as living and thus *autonomous*.

This « darwinian natural selection » mechanism would indeed seem to have this fundamental virtue in terms of *autonomy* of not requiring the formation of *categories of coherent actions* that are unfeasible for a solitary due to of the 'indistinguishability theorem'.

This mechanism is called « evolutionary robotics » and would involve not only one exploration robot but several of them. To this end, a number of robots should first be sent to martian ground. Upon returning from their first mission, those robots that would not have been damaged could then 'marry' by combining their artificial 'genes'. Hence the emergence of a new reduced flotilla of 'son' robots having inherited the experience fortuitously acquired by their 'parents'. In practice, these artificial 'genes' could be constituted by the 'synaptic weights' of the artificial neural networks that would make up the controllers of the robot fleet.

After several generations of robots having explored the martian ground, there should therefore theoretically exist according to the theory of « natural selection » one or more robots that would have learned without the help of a technician to avoid all hot objects on the martian ground that could have destroyed them.

One or more exploration robots would thus have become spontaneously *autonomous*, which would certainly go against our previous statement that the *autonomy* of a solitary exploration robot was physically infeasible because it could only react randomly to the perception of objects in its environment.

Evolutionary robotics experiments carried out in several laboratories show that this « natural selection » mechanism is quite operative. But the rational analysis of the *functional* duplication mechanism that allows the birth of 'son' robots that have inherited from their 'parents' through the mixing of the 'genes' of robots that would have returned unscathed from their explorations, shows that this mechanism is in reality physically unfeasible due to the 'indistinguishability theorem'. Like the unfeasible formation of *coherent categories of actions* that would found the *autonomy* of a solitary exploration robot.

If these laboratory experiments of « evolutionary robotics » are nevertheless quite convincing, it is because technicians have necessarily intervened in a very oriented way in the implementation of algorithms related to the 'wedding' process of robots defined by their genes.

As a technician must necessarily be involved in the proper installation of electrical connections between the temperature sensor and the relay of a thermostat.

In the absence of any technician, the mechanism of « darwinian natural selection » applied to robotics would therefore have no expected ability to render *autonomous* one or more robots of the martian ground exploration flotilla.

The functional analysis of the « darwinian natural selection » process, which we know to be constantly at work in nature by creating a multitude of new species, shows that it actually only selecting among the various possible *forms* of already *autonomous* material systems, those that are best adapted to survive environmental constraints. This process of selecting *forms*, which uses only *imprint* reproduction mechanisms and not *functional* ones, is perfectly lawful in view of the 'indistinguishability theorem'.

Consciousness

In short, the realization of an *autonomous*, artificially *living* robot, which is the goal we had set ourselves, is therefore physically impossible according to the 'indistinguishability theorem'. At best, we can only build an *automaton* that will only obey the limited orders that its creator will have implanted in its memory.

That an *autonomous* robot, artificially alive, is finally physically impossible is undoubtedly somewhat disappointing for the proponents of a *strong* artificial intelligence where the various functionalities of the living, and in particular the *consciousness*, must all be reduced to algorithms. But what is ultimately quite paradoxical in the strong sense of the term is that there are many physico-chemical structures on Earth that are naturally *autonomous*, they are the living beings !

Let us consider a technician, a *living* physico-chemical structure, who observes the movements of the mercury column of a thermometer that he holds in his hands decides to carefully move away from all very hot objects that could burn him and thereby shorten his life.

The success of this operation of systematic avoidance of hot objects therefore implies that this technician must be specifically equipped with an « operator » whose essential virtue is to be able to *differentiate* between the different positions of the mercury meniscus of the thermometer which are physically indistinguishable between them according to the 'indistinguishability theorem'.

Experimentally, in the ultimate analysis, we note that it is the 'pleasure' felt by the technician in the success of the operation to avoid hot objects that could destroy him, that would be at the origin of this singular ability that the living possess to *differentiate* objects of the world that are fundamentally *indistinguishable*. It is also for the 'pleasure' of the technician who has taken a lot of trouble to build an exploration robot, that it is therefore imperative that this robot subsequently turn away from any type of hot object that would inevitably cause its destruction.

The 'pleasure' that the technician currently feels in preventing the robot he built with great difficulty from colliding with hot objects that are on its path, has its origin in his past experience of being alive. For instance, it was by visiting a foundry in which the technician began to get very hot, which seriously bothered him, that the particular technical solution {*turn away* from the observed incandescent object} created *by chance* by his brain-computer from sensor data, was spontaneously activated

The fast moving away from the hot object had the beneficial effect of rapidly reducing his thermal discomfort by lowering his body temperature and thereby the concomitant appearance of a 'pleasure' who had in that very moment forever *marked* the action {*to turn away* from the observed object} which had then been memorized in his brain-computer.

Hence the current reaction of the technician who, for his 'pleasure', tries to prevent the robot from colliding with hot objects on its path, as in his past experience.

Empirically, the *consciousness* endowed with the extraordinarily pregnant *sensitive qualities* of 'pleasure' and 'pain' would be the intended operator because it has the unique power to differentiate for 'its pleasure to be alive' the objects of the world which, fundamentally, are physically *indistinguishable*.

This mode of action of consciousness, which would be reduced to choosing particular technical solutions among those created blindly by our brain as a computer, is in perfect agreement with the

paradoxical results of the experiments of the neurobiologist Benjamin Libet : « consciousness vetoes solutions previously developed some 500ms earlier by the brain-computer. »[2]

As a corollary, the operability of 'mental objects', which result from the multifaceted utilitarian divisions of a universe that we perceive through the sensors we are equipped with, would result from the perennial choices made by *sensitive qualities*. As such, the actions that flow from our thoughts would all be fundamentally 'irrational'. In the sense that the *actions* that would be updated among all those possible resulting from physical interactions, as for them required, would in no case result from *logical* operations based on the laws of physics. Despite the fact that these living beings necessarily possess *rational* knowledge about the world in order to be able to act on it in an effective way to ensure the *durability* of their structure.

Essential property, the *consciousness* that the technician naturally possesses, and more generally all living beings, is that considering the 'indistinguishability theorem', *irreducible* to any neuronal physicochemical interaction of the brain. In other words, it would be of a *non-physical* nature as it is *non-measurable*. If *consciousness* were measurable, it could be defined by complementary dual attributes such as hot/cold, white/black,... and the computation of predicates (or descriptors) applied to the new system {sensor + consciousness} would then show that adding this *consciousness* to the sensor system whose states were originally *indistinguishable* would only increase the number of states perceived by the sensor system without reducing the *indistinguishability* of its states. Consciousness would therefore be of a *non-physical* nature.

This hypothesis of the non-materiality of consciousness is not physically irrelevant as one might at first think, based on the assumption commonly adopted by the scientific community that the 'material dimension' is the only possible 'dimension' of the universe.

The 'material dimension' is not an 'object' that can be observed through measurements. It is only a *concept*, and therefore results from a process of *coherent categorizations* which, as we have seen, implies the existence of *consciousness*, a « first operator », which is necessarily *irreducible* to any physical interaction because of the 'indistinguishability theorem'. As such, this hypothesis of the *non-materiality* of consciousness would therefore be legitimate.

Thus, the *states of consciousness* that open us to the colourful and sensitive perception of a universe otherwise closed on itself, would themselves be fundamentally *irreducible* to the mental states of our brain contrary to the theory of 'mind-brain identity' (IT).

These *states of consciousness* would therefore not be the fruit in the physico-chemical sense of the term of specific neural states, but only induced by these neural states. Consciousness is not like the memory of a computer that results from the combination of a multitude of silicon-based electronic components (semiconductors).

As such, *consciousness* could not be reduced to an algorithm as stated by the neuroscientist Stanislas Dehaene (Collège de France) in his book 'Consciousness and the brain : deciphering How the brain codes our thoughts' where it is about a code that describes the location of the different neurons that are activated when the subject says to be *conscious*.

In the experiment that has been carried out, it was only a question of seeking if there were some correlations between the statements made by a subject about for instance a 'pain' he was feeling, and the neural activity of certain regions of his brain, which was being viewed using brain imaging techniques or electrodes implanted in his cortex. But these measurements did not provide in any way access to the *consciousness* itself, to its operative capacity that founds the *living*. Theses researches does not invalidate in any way the thesis that we support regarding the irreducibility of *consciousness* to any physical interaction.

Conclusion

Ignored by the current dominant thinking on living things, just as the *transhumanist* thinking which is reasoning only in mechanical and computational terms, the *consciousness* that opens us to the world with its perfumes, its resplendent aurora, its pleasures but also its pains, is nevertheless proving to be

[2] Libet Benjamin – *Unconscious cerebral initiative and the role of conscious will in voluntary action. Neurophysiology of Consciousness,* pp. 269-306 – Contemporary Neuroscientists 1993.

the key to the existence on Earth of living beings and their cognitions. These are the *consciousness* that living beings are endowed with that determine how we act on the world to ensure our durability, that is, to be alive.

Far from being an epiphenomenon, *consciousness* would therefore be an essential term, in other words a « first operator » without which no *living* being would have appeared on Earth.

Even the brains of today's living beings, from the most elementary like that of bacteria composed of some proteins associated in networks, to the most complex as that of man, would only be physico-chemical machines. Machines which, in the absence of *consciousness* who carry out appropriate choices among the possible technical solutions created by these machines, would only *randomly* combine information from different sensors. A process that would be antinomic with the formation of *coherent categories of actions* that underlie *living* organisms.

For a human being to be more clever, he would fundamentally have to be able to create a large number of new *concepts* which, by definition, are *coherent categorizations* of the world objects that a living being perceives with his sensors in order to act in an appropriate way that ensures his durability. For example *categories of coherent actions* or *concepts* {'flee' for all temperatures above 30° C} and {'moving forward' for all temperatures below 30° C} that are assigned to the exploration robot and which should allow it to avoid 'cleverly' all hot objects.

But to create new *coherent categories* it would not be enough, for example, to simply increase the capacity of one's memory as suggested by the *transhumanist* thesis. It would also be necessary to fully control the operative properties of the *consciousness*. Otherwise, the information processed by the brain, as a computer, could only be loaded into the memory in a *random* way, since this information would then be strictly *indistinguishable* with regard to the 'indistinguishability theorem', and therefore without any further possibility of ordered readings.

But we have shown that due to the fact that consciousness has the physically paradoxical capacity to make choices among objects of the world that are physically *indistinguishable*, logically implies that this *consciousness* is strictly irreducible to any physical process. And this is how *consciousness* can not result however complex it may be.

One cannot thereby repair or create a consciousness as one builds a computer or graft a piece of reconstituted heart tissue onto a failing heart. Even if the progress of science is such that one day we can build adequate physico-chemical structures which, experimentally, are found to have the capacity to induce *consciousness*, the fact remains that the fundamental irreducibility of *consciousness* to any physical interaction one will probably never be able to control the paradoxical capacity of consciousness to select 'for its pleasure' technical solutions that are fundamentally *indistinguishable*.

The *transhumanist* thesis that states that the only increase in the computing power of our brains should allow unlimited increase in the intelligence of human beings is therefore unfounded

The only thing that is physically possible is to modify or even suppress the *induction* of certain 'sensitive qualities' or constituents of *consciousness*, such as 'pleasure', 'pain', 'colour', 'sound',... But without mastering the decision-making capacity of the *consciousness* that founds the state of life and therefore the existence of all the objects that are built by living beings.

As such, the *transhumanists'* proposition to build an artificial brain able of creating, as we naturally know how to do, from only mechanical, computing, chemical, interactions between technical components, would be totally unfounded. All creation implies a *consciousness*.

As for the hypothetical transfer of our *mind* to a supercomputer by simply copying the different states of activation or non-activation of brain neurons assimilated to a computer machine as suggested by the *transhumanist* thesis, i. e. by ignoring the existence of *consciousness*, is just as impossible. We would only build a super-automaton like the 'Automaton' of neurobiologist Wilder Penfield, that is to say a human being deprived of *consciousness*, therefore of any *sensitivity*, because of a major dysfunction or surgery, and thus totally lost the ability to create, to adapt to an environment different from the one he had known when he was in good health.