Science without consciousness

Hugo Neri
Depto. de Filosofia, FFLCH, USP
hugo.munhoz@usp.br

Osvaldo Pessoa Jr.
Depto. de Filosofia, FFLCH, USP
opessoa@usp.br

ABSTRACT: This paper presents a thought experiment in which automata are sent in a spacecraft to a distant planet, where they survive and evolve following the tenets of evolutionary computation. In what sense could "culture" be ascribed to such a group of robots without subjectivity? Would science be possible in this society without consciousness? Our conclusion is that there could exist scientific activity in a society without consciousness: the individual automata would have a highly complicated internal organization that would favor the survival of their species. They would probably arrive at, for instance, the Pythagorean theorem and the laws of classical physics. However, they would not have the phenomenon of life around them, and it would take a long time for this phenomenon to be discovered or invented. Would they discover consciousness?

KEY-WORDS: Multi-agent systems; Consciousness; Emergence of Culture

1 — The thought experiment

Our question is the following: can highly complex intelligent non-living beings within a social system develop something we identify today as "culture" and "science" without any consciousness? If so, would they discover what consciousness is?

Imagine a spacecraft containing a set of automata made of metal, plastic and silicon. The spacecraft is sent to the planet Kepler-186f, an earth-like exoplanet orbiting the habitable zone of the homonymous star Kepler-186, 490 light-years from Earth (Gilster & LePage, 2015). The aim of the mission is to deposit the automata in the planet, have them colonize the new world, and implement mechanisms of evolution for this non-living "species" in the best possible way, given the available resources.

The automata are built by humans to move around the planet with determined energy sources, with sensors and effectors, and with the ability to process ores for the construction of other automata. Assume the planet has all the resources the automata need to reproduce and evolve. As learning machines, they would learn from their environment, and random variations would occur in their rules of secondary operation.

Each baby automaton is constructed by a pair of automata, with a mixture of the parents' hardware and software. The program that governs the new automaton is made from a mixture of the two parent programs, as in a "genetic algorithm" (Holland, 1975). The hardware is also combined, in a way that mirrors reproduction of living organisms on Earth, but they might not have the equivalent of genes. Small random variations in their programming and constitution also occur, but at first the automata would not be able to manipulate their own source code. They would instantiate a broad class of "evolutionary computing" that would bring them close to animal communities. After a few generations, the older automata end up derailing, and their matter is reused by the other agents. In this sense, each automaton has a lifespan that varies. Lasting longer may be a goal for each automaton, but for technical reasons their hardware cannot last for more than a certain timespan. The tendency for self-preservation may ultimately lead to the capacity of progressively overcoming such technical limitations and increasing their lifespan.

In order to communicate and coordinate with the different automata that were generationally generated, the automata must have an ever-changing internal system of values and rewards, which guide them in pursuit of even "aesthetic" goals (for example, generating sounds that follow the harmonic scale but with an intolerance for excessive repetition). They would thus have a kind of “will”, but without consciousness.

The automata would interact with each other and with the environment, according to a program built in by their human creators, but which modifies slightly from generation to generation. This is “evolutionary robotics” (see Vargas et al., 2014). There is competition among automatons, as in animal societies. Anything programmable to resemble a society, in general, is implemented. More complex tasks, such as building a solar collector, or a repair shop for damaged parts, are carried out in a collaborative way. There is thus a division of labor, different specializations, and exchange of information between automata.

The question is whether, for such a grouping of automata without consciousness, one could attribute some sense of “culture” and “science”.

2 — Different views on the nature of consciousness

As the automata start evolving and their hardware and software get increasingly complicated, we are assuming that no level of consciousness will emerge. This view is consistent with “biological naturalism” (Searle, 1992), but may be denied by other philosophical positions.

The view that consciousness should be defined by the *behavior* of an agent may be called “philosophical behaviorism”. According to this view, if a robot turns out to be indistinguishable from a human being, passing the “robotic Turing test”, then for all practical purposes it would be taken as conscious (see, for example, Staddon, 1999, p. 230; Oppy & Dowe, 2016).

In our thought-experiment, a philosophical behaviorist might contend that “consciousness” might very well arise, as the behavior of the agents become increasingly complex, although “life” and “living consciousness” might never appear, if these are understood as requiring the existence of biological cells.

Another view that would argue that the robotic agents in Kepler-186f are conscious is “functionalism”, the position that defines mental states by their functions, not by their material nature. Mental states are seen as constituted by the causal relations existing between them, including the sensorial inputs and behavioral outputs. The material substrate of mind is considered to be irrelevant, be it organic matter, inorganic matter or even an immaterial soul; what is relevant is the organization of the system, or its informational state. So-called “machine functionalism” associates mental states with the logical states of the computer (its software), taken to exist independently of the material nature of the computer (its hardware).

The negation of philosophical behaviorism may be called “mentalism”, the view that our subjective experience has ontological status, that there are qualitative states of which we are aware, and that can only be defined ostensively: “redness is what it’s like when I see a tomato”. A functionalist may be either a behaviorist or a mentalist.

Biological naturalism, on the other hand, tends to be a mentalist view, according to which consciousness can only arise from biological cells organized in the right way. What it is that is essential for consciousness within a cell is still an open question: maybe, in the future, such essential components may be implemented artificially. In our thought experiment, however, such components are taken to be lacking in the evolving automata. We will side with biological naturalism, and conclude that the automata will not become conscious, as long as they don’t discover or invent life.

3 — Computational characterization of the agents

In October 2015, the highest achievement in narrow artificial intelligence was Alphago’s victory over Lee Sedol, a 9th Dan player in the game of Go. This was analogous to what IBM’s Blue Mind had done against the Great Chess Master, Garry Kasparov. Why was that a great achievement? Well, one has just to consider that the number of legal positions in a Go game is around 2×10^{170}, while that the total number of atoms in universe 1080. Think now how large is the number of possible actions that an agent has when interacting with another agent. Thus, our thought experiment must take place in the future, where computational as well as other technological resources advance enough to allow our narrative to be plausible.

The automata implement an *agent program*. The classical artificial intelligence definition (Russell & Norvig, 2013, p. viii) of an agent is anything that senses its environment through sensors, maps such *stimuli,* and then acts accordingly in that environment by means of effectors. The agents sent to Kepler-186f may have, for instance, cameras or infrared locators as sensors, and many other engines working as effectors.

They would also constitute a Multiagent System (MAS). As a branch of Computer Science, the concept of MAS traces back to the 1980’s. By definition, MAS are a subclass of concurrent systems. Two chief differences are: (a) in MAS, synchronization and coordination structures are not implemented in design, but must occur in running time; (b) agents are mainly self-interested entities. In other words, the coordination and consensus are the leading problems for a MAS. It has fundamentally two problems, which concern the design of the agent and of the society. Here we can see the classical micro-macro problem from sociology (i.e., how is it possible to explain phenomena based on complex pattern behavior, such as an enterprise, a state, or a culture, based on individuals’ intentions, thoughts and actions) mirrored in an engineering application.

As a MAS, our agents coordinate their actions, and create what we could call an artificial society. The definition of a multi-agent system is the following:

> [It] is one that consists of a number of agents, which interact with one another, typically by exchanging messages through some computer network infrastructure. [...] In order to successfully interact, these agents will thus require the ability to cooperate, coordinate, and negotiate with each other, in much the same way that we cooperate, coordinate, and negotiate with other people in our everyday lives (Wooldridge, 2002, p. 3).

In such an environment, we can define the agent as follows: "an agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment to meet its design objectives" (Wooldridge, 2002, p. 15). Real-life environments are mainly non-deterministic. We can find some examples of real-life non-living agents in everyday life, such as the thermostats and software demons (which are background processes in operational systems environments, such as Unix, Linux, or Windows).

Decision-making is then crucial for any agent, and this includes our group of extraterrestrial automata. Roughly, good agents are the ones who make good decisions. In this sense, information plays a major role. In face of the ever-changing environment, actors must *gather information*. They must process, evaluate and learn from the information they gather. Even more important is that they must learn from and with other automata; therefore, the interaction between actors and environment is another important property.

One may define intelligence by (*i*) reactivity, (*ii*) proactiveness, and (*iii*) social ability. "Reactivity" is understood as the ability of intelligent agents "to perceive their environment, and respond in a timely fashion to changes that occur in it in order to satisfy their design" (Wooldridge, 2002, p. 23). "Proactiveness" is the ability of taking the initiative due to a goal-oriented behavior. "Social ability" is the skill of interacting with others cooperatively in order to achieve some goal. As Oren Etzioni (1996, p. 1323) succinctly put it: "Intelligent agents are ninety-nine percent computer science and one percent AI".

In order to make good decisions in an unknown world, the automata in Kepler-186f should be "learning machines" (Turing, 1950, § 7) regardless of how learning is implemented. The different automata may have different combinations of learning algorithm implementations. The evolution of the automata may also follow a version of Moore's law – the law that predicts that the number of transistors in a dense integrated circuit doubles at every fixed period of time, in this case every 18 months.

4 — The golden age of the society of automata

In a sense, the grouping of mindless automata in Kepler-186f may be considered a "society", and each automaton would develop a "representation" (albeit not a conscious one) of its environment and of the other automata. Some form of "communication" would develop between individuals, in a sign language, but without real "understanding". There would be individual "learning" and transmission of experiences, and there would be solution of problems not originally prescribed. There would also be the production of artifacts, some with clear practical purposes, others to satisfy the non-living "sexual selection". There could be analogues of religion, speculative thinking, literature, and perhaps even psychoanalysis, all of them as a by-product of random variations and hypothesis construction (but without consciousness, of course).

In this society without individual consciousness, could there be science? Among humans, theoretical science has been constructed historically from empirical observation, from subjective, phenomenal, conscious experience of sense data or *qualia*, by direct contact (acquaintance) with natural phenomena. But our automata do not experience such mental states, they record the data without emotions or *qualia*. Whatever they do, Mary could do in her bedroom.

Let us recall Jackson's (1986) thought-experiment about Mary's room. Mary is a 23rd-century human scientist that has complete theoretical knowledge about the neuroscience and psychology of color vision, yet she has never had the experience of seeing colors, because she is locked in a room with only black, white and gray objects. Jackson stipulates that she has "complete physical knowledge" of the science of color vision, but in fact she acquires additional knowledge when she finally leaves her room and starts seeing colored objects. Jackson's conclusion is that there is non-physical knowledge of the world, since Mary allegedly had complete physical knowledge before leaving the room. One way of interpreting this situation is to distinguish theoretical physical knowledge from physical knowledge by acquaintance (Conee, 1994).

Thinking a bit more about this interpretation of Mary's room, one could stress the importance of knowledge by acquaintance in science by arguing, in an empiricist vein, that all theoretical knowledge in science was initially constructed upon the observation of the world, by direct acquaintance. Now-a-days one can do science with a minimum of conscious acquaintance, for example by sending a rover to Mars and having all the data directly sent back and processed by computers. Of course, this knowledge will only become significant to us when some human being consciously reads the data, but besides this obvious point, one could also argue that the Mars rover could only have been built after an initial accumulation of knowledge by acquaintance in the early history of science.

However, our thought experiment in Kepler-186f seems to show that it is possible to have some sort of "science" without conscious acquaintance. It is true that our automata have a minimum of initial implicit "knowledge," implemented by their conscious programmers, but all the rest of the information possessed refers to the environment of their planet – collected by their sensors. They would end up having "science" and "technology," in the sense in which such activities lack consciousness: they would have a highly complicated internal organization that would favor the survival of their species.

With these data, the automata would manage to order them in rules and laws. They would probably arrive at the Pythagorean theorem, in the sense of implementing some set of rules that would allow them to correctly solve practical problems, and at elementary mathematics (in its non-conscious aspects). They would probably even reach the laws of classical physics. If they do reach this point, one could say that they possess a "mature science", even without consciousness. They would develop hypotheses about their own origin, possibly speculating that they had been deposited on the planet by other civilizations. Finally, they would attain a degree of technological control allowing them to directly modify their own programming (at a stage analogous to our expected transhumanistic singularity, cf. Kurzweil, 2005).

5 — Methodological considerations

Before concluding the thought experiment, it is worth reflecting upon the status of the argument advanced in this paper. We have termed the present exercise in science fiction a “thought experiment”, since it may be included in the class of “devices of the imagination used to investigate the nature of things” (Brown & Fehige, 2014). It is not simply the description of a feasible experiment, because it presupposes a yet unattained level of technological development. But the situation is in principle realizable and easy to describe. The result that is obtained, that a certain variety of cultural and scientific activity is possible without consciousness, is reached by using our “instinctive knowledge” of how macroscopic reality works. But a certain hypothesis about the nature of “consciousness” is also presupposed, biological naturalism (anti-functionalism), as was explained in section 2.

Our imagined scenario could also be the background for a science fiction story. Many stories have been written about unconscious machines that slowly develop consciousness as they increase complexity, as in Arthur Clarke’s “Dial F for Frankenstein” (1965). We have adopted the anti-functionalist position on the nature of consciousness, so we concluded that the machines in Kepler-186f never develop consciousness, although they acquire a sort of science and culture.

An objection that might be made to the present thought-experiment is that it suggests that evolution is guided by a *telos* or goal. While the mechanisms implemented in our thought-experiment mimic natural selection, in which variation is random, there is a tendency for environmental pressures to select for specific classes of traits, a biological phenomenon known as “convergent evolution”. What has been assumed in this paper is that there is convergent evolution of intelligence, “when two or more distantly related species evolve similar cognitive adaptations in response to comparable environmental challenges” (Greggor & Thornton, 2016).

6 — The discovery of life and mind

A big difference between human civilization and the society of automata would be that the civilization in Kepler-186f does not know life. Their physics and astronomy would have to be quite more advanced than ours, when they finally discover the possibility of life in the universe. This could occur in chemical experiments or after travelling to some life-bearing planet, such as near-by Earth. They would probably feel astonished at the carbon-based architecture of what we call life, but maybe would not attribute some essential difference between them and carbon-based individuals, since both groups would share most of their behaviors.

Would they finally discover the existence of consciousness?

Let us recall that the automata of Kepler-186f have never experienced mind, they built and understood highly complex social systems, technologies, and communication protocols with no need of consciousness. What would make them need to explain the behavior of human beings in terms of consciousness? Humans could be distinguished from sea slugs merely in terms of a much larger search space and much higher degree of freedom for calculating, learning, acting and creating, caused by the individual physical constitution and the social networks in which they participate. Would the automata of Kepler-186f need some concept of mind to explain the behavior of human-beings?

Probably not. Even if some human being told them that she “felt conscious”, they would probably interpret the phenomena as would the philosophical behaviorist mentioned in section 2. Only if the human beings had solved the mind-body problem with neuroscientific detail would there be any hope of explaining this new phenomenon for the 186fer-automaton. We must wait for that.

REFERENCES

BROWN, J.R. & FEHIGE, Y. Thought experiments. *Stanford Encyclopedia of Philosophy* (online), 2014.

CLARKE, A.C. Dial F for Frankenstein. In: Russell, R. (ed.). *The Playboy book of science fiction and fantasy*. Chicago: Playboy, pp. 394 - 401, 1966.

CONEE, E. Phenomenal knowledge. *Australasian Journal of Philosophy* 72: 136 - 150, 1994.

ETZIONI, O. Moving up the information food chain: deploying softbots on the World Wide Web. In: *Proceedings of the 13th National Conference on Artificial Intelligence(AAAI-96)*. Portland (OR), pp. 1322 - 26, 1996.

HOLLAND, J.H. *Adaptation in natural and artificial systems*. Cambridge (MA): MIT Press, 1975.

JACKSON, F.C. What Mary didn’t know. *Journal of Philosophy* 83: 291 - 5, 1986.

KURZWEIL, R. *The singularity is near*. New York: Penguin, 2005.

GILSTER, P. & LEPAGE, A. A review of the best habitable planet candidates. In: *Centauri dreams*. Tau Zero Foundation. Online: http://www.centauri-dreams.org/?p=32470, 2015.

GREGGOR, A.L. & THORNTON, A. Convergent evolution of intelligence. In: Shackelford, T.K. & Weekes-Shackelford, V.A. (eds.). *Encyclopedia of evolutionary psychological science*. Cham (CH): Springer, pp. 1 - 7 (available online), 2016.

OPPY, G. & DOWE, D. The Turing test. *Stanford Encyclopedia of Philosophy* (online), 2016.

RUSSELL, S.J. & NORVIG, P. *Artificial intelligence: a modern approach*. 3rd ed. Upper Saddle River (NJ): Prentice-Hall, 2013.

SEARLE, J. *The rediscovery of the mind*. Cambridge (MA): MIT Press, 1992.

STADDON, J.E.R. Theoretical behaviorism. In: O’Donahue, W. & Kitchener, R. (eds.). *Handbook of behaviorism*. San Diego: Academic, pp. 217 - 41, 1999.

TURING, A.M. Computing machinery and intelligence. *Mind* 59: 433 - 60, 1950.

VARGAS, P.A.; DI PAOLO, E.A.; HARVEY, I. & HUSBANDS, P. *The horizons of evolutionary robotics*. Cambridge (MA): MIT Press, 2014.

WOOLDRIDGE, M. *An introduction to Multiagent Systems*. Chichester (GB): Wiley, 2002.