Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 2 (2020) 183–198

DOI: 10.1142/S2705078520710022

World Scientific
www.worldscientific.com

# Conscious Machines: A Possibility? If So, How?

Riccardo Manzotti

*Department of Business, Law, Economics and Consumer Behavior*
*IULM University, Milan, Italy*
*riccardo.manzotti@iulm.it*

Antonio Chella*

*Department of Engineering*
*University of Palermo & ICAR-CNR*
*Palermo, Italy*
*antonio.chella@unipa.it*

Published 24 July 2020

The scope of the paper is to encourage scientists and engineering to avoid to do what Einstein pointed out as being the hallmark of folly. Machine consciousness scholars must be brave enough to step out of the beaten path. There must be some big recurrent conceptual mistakes that prevent science and technology from addressing machine consciousness.

*Keywords*: Machine Consciousness; Artificial Consciousness; Robot Consciousness.

## 1. Background

According to a popular belief, Albert Einstein once said that repeating the same actions and yet expecting a different account is akin to madness. While it is uncertain whether he said exactly anything like that he surely wrote that *In the interest of science it is necessary over and over again to engage in the critique of fundamental concepts [. . .] This becomes evident especially in those situations involving development of ideas in which the consistent use of the traditional fundamental concepts leads us to paradoxes difficult to resolve* [Einstein, 1954, p. 14]. Here is a good definition of the state of the studies about the replication of consciousness in machines. Ever since the onset of AI — assuming the Dartmouth summer school as the conventional starting milestone [McCarthy *et al.*, 1955] — there has been a series of repeated attempts at replicating the conscious mind using machines. Whenever the available technology has promised to achieve some unmatched mental skill, hopes were reawakened.

*Corresponding author.

AI is now entering into a period of unprecedented technological development [Harari, 2018] whose outreach is still mostly unknown. Is consciousness in machines among the possible outcomes? It's a fact that the progress in the field of AI has not been a linear one and it has had its ups and down, technical developments have often led to unrealistic expectations that have backfired and led to a series of AI springs and winters [Dreyfus, 1972, 1992; Searle, 1992; Yasnitsky, 2020]. Yet, let's suppose that the necessary technological ingredients are available, is the scientific community ready to implement and design an implementable model of consciousness?

In this paper, we want to address the current landscape in machine consciousness. In the past, we already dubbed conventional strategies of Good Old Fashioned Artificial Consciousness (GOFAC), [Manzotti and Chella, 2018]. While these strategies have been based on respected approaches to AI, they have not been successful.

So far, there have been various attitudes. First, it is fair to admit that nobody has a clue as to what a theory of consciousness should be. Despite years of attempts at finding a solution to the problem, we can candidly admit that we are still at square one. We have learned a lot about issues that are, reasonably, somewhat connected to consciousness, but nothing we have learned so far has led to consciousness. For instance, we have learned a lot about face recognition. However, do we have any reason to conclude that what we have learned so far about face recognition, despite its remarkable applications to defend and/or threaten our privacy and security, is going to shed light on the conscious experience we have when we see a face? Nobody expects our cell phones to be conscious even if they very reliably unlock our phones in response to our faces. Likewise, nobody expects DeepMind AlphaGo to be conscious of its moves, despite being able to defeat the world's number one human Go player KeJie [Silver *et al.*, 2017]. While the same cognitive skills in humans are associated with consciousness, there is so far neither evidence nor theoretical necessity as to why the implementation of AI should lead to anything like consciousness.

Different kinds of sciences call for different types of theories. Physicists, for example, are searching for a "grand unified theory" that integrates gravity, electromagnetism and the strong and weak nuclear forces into a homogeneous set of equations. Whether or not they will get there, they have made considerable progress, in part because they know what they are looking for. AI cannot hope for that kind of theory. Intelligence and cognition aren't elegant the way physics appears to be. Cognition is the outcome of historical accidents, with agents solving problems based on circumstances that lead them down one solution rather than another.

But cognitive complexity is only part of the challenge in figuring out what kind of theory of consciousness we're seeking. What we are looking for is a bridge, some way of connecting two different and mostly separate domains — those of science and individual experience. Such fields have been defined in entirely different ways from the onset of science itself. In 1958, the great physicist Edwin Schrödinger wrote that "What particular properties distinguish these brain processes and enable them to

produce [consciousness]? Can we guess which material processes have this power, which not? Or simpler: What kind of material process is directly associated with consciousness?" [Schrödinger, 1958, p. 93]. And, of course, Schrödinger's echoed almost verbatim the famous Thomas Huxley's quotation [Huxley, 1863]. We are still there.

This vagueness of the target has had apparent consequences on all attempts to reproduce consciousness employing machines. There seems to be a need for a conceptual and scientific breakthrough that has not yet been achieved. In this regard, Gary Marcus recently wrote that "Such bridges don't come easily or often, maybe once in a generation, but when they do arrive, they can change everything. An example is the discovery of DNA, which allowed us to understand how genetic information could be represented and replicated in a physical structure. In one stroke, this bridge transformed biology from a mystery — in which the physical basis of life was almost entirely unknown — into a tractable if challenging set of problems, such as sequencing genes, working out the proteins that they encode and discerning the circumstances that govern their distribution in the body. Neuroscience awaits a similar breakthrough. We know that there must be some lawful relation between assemblies of neurons and the elements of thought, but we are currently at a loss to describe those laws." [Marcus, 2014]. Yet, the temptation to build a conscious machine is too great to resist.

We are aware that such suggestions may be seen by many as not being prudent enough. The proliferation of gurus and scientifically unsupported views has led to a widespread skepticism in the field [Lau and Michel, 2019; Michel *et al.*, 2018, 2019]. Yet, the area is young, and it is understandable and indeed desirable that new ideas are proposed and discussed. In fact, it has been argued with a mathematical model that in a new field, such is the case of consciousness studies [Akerlof and Michaillat, 2018]:

> science may converge to the worse paradigm if few scientists initially believe in the better paradigm. [. . .] That the worse paradigm may prevail at all may be surprising since neither is it more fruitful for research nor do its adherents have greater bias. The worse paradigm may prevail nonetheless because once it has many adherents, a tenure candidate is highly likely to be evaluated by a worse-paradigm scientist. [. . .] the number of tenured scientists believing in the worse paradigm grows faster than the number of tenured scientists believing in the better paradigm: Science moves away from the truth.

In this paper, we will map the current theoretical landscape in machine consciousness. Thus, we will encourage scientists and engineering to avoid to do what Einstein pointed out as being the hallmark of folly. We must be brave enough to step out of the beaten path. There must be some big recurrent conceptual mistakes that prevent science and technology from addressing machine consciousness.

## 2. A Caveat

Before proceeding further, a necessary caveat is to be mentioned — the expression *machine consciousness* (and occasionally *artificial consciousness*) should not imply that consciousness is a property of machine themselves. Neither do we mean that there is a difference between *machine* consciousness and *natural* consciousness. The reason is that there is so far no conclusive evidence as to what is conscious, and thus, we do not know whether consciousness is an attribute of something going on inside a machine or even a characteristic of the device. We have made this point in the past, but it is always important to restate it [Chella and Manzotti, 2009a,b, 2012; Manzotti and Chella, 2018]. The expression *machine consciousness* refers to any attempt to achieve something akin to what we call consciousness by using an artificial structure, be it a computer, a software running inside a computer, a robot, or any combinations of them. This is to say that it might turn out that neither brains nor machines are *conscious*, although they might be one of the necessary ingredients to allow consciousness to occur. While it might be evident to many that a brain is conscious, it is not sure that it is the case, unless a method to measure consciousness had been devised, which has not. It would then be more cautious about using expressions such as "achieving consciousness through machines," but it might sound too repetitive.

Why are the notion of a conscious machine and that of a conscious brain questionable? The problem is that, if one is a physicalist, consciousness cannot be added to a physical system as an additional property. So, if one assumed that there might be, say, a conscious neural network, one could always, in principle, conceive of the same neural network without consciousness. Consider applying the notion of the philosophical zombie applied to machines and neural machinery [Chalmers, 1996; Kirk, 1974] — if we define consciousness as an additional property to the physical ones, it will always be conceivable to conceive of the same physical structure unless consciousness was yet another tangible property. Still, by doing so, we will deny the premises we started from.

For example, consider an electronic gate with currency flowing through it. Such an electronic component can be exhaustively described in terms of its causal properties. This is indeed what we are going to find in the manufacturer's spec sheet. Suppose now that the component is inserted inside a "conscious machine." Would it be any different? Of course not. The gate will continue to have the same causal properties its manufacture designed it for. It cannot be any different because it is now part of a conscious machine. The causal properties of the gate would be the same regardless of whether the gate is part of a conscious machine or not. The same rationale might be applied to any subset of the machine up to the device itself. While there have been frequent claims as to the emergence of top-down causation, there has so far been no convincing evidence about it. As far as we know, all causal powers are drained by a machine's components that are what they are. This is not to say that consciousness is not a real phenomenon, yet that there are convincing arguments against the

possibility that consciousness is an attribute of the machine or their subsets. Is this an argument against internalism? Yes, it is.

Another issue that we want to stress is that the software cannot become conscious. The software is an abstract layer of a description of a more mundane and complicated state of things, which is the physical configuration of a machine. The software is a high-level description that allows us to quickly rearrange the physical structure of a processor to make it behave according to a desired causal pattern. If we looked for our symbolic languages inside a machine, we would not find anything. The belief in the existence of layers inside a computer is only a widespread modern myth. Inside computers, there are only tiny silicon-based micro transistors acting as switches. Software is a handy way, very convenient for humans, to translate a high-level description into one of the many causally similar patterns.

This is why we believe that the advent of *conscious software* is hardly possible. One may counter argue that the above arguments also rule out the possibility of *conscious brains*. Yes, we admit that and we believe that this is not a counter-argument, but rather a reasonable conclusion that should be taken in due account.

## 3. Common Pitfalls

Notwithstanding the previous worries, the reality of consciousness in biological beings remains and, because of that the possibility to replicate its occurrence by means of machines. The above worries should not be taken as a cause for despair but rather as an encouragement to look beyond current conceptual frameworks (and theoretical fences). In order to do that, we would like to list the most common conceptual drawbacks that have so far hampered the technological reproduction of consciousness. While there are no definite answers, it is possible to mention a few approaches that in the current scientific world are very likely the source of much confusion. They are listed here in no particular order. Of course, the following positions are contrasted with the issue of consciousness. So, our criticism is not directed toward these positions as such, but toward such positions in regard with consciousness.

### 3.1. *Computationalism*

In the field of consciousness research, computationalism is the tendency to deal with either information or computation as they were additional substance that is somewhat generate by a computing machine. This way to frame the issue may seem very crude and approximate and of course no serious information theorist would endorse it [Capurro and Hjørland, 2005; Qvortrup, 1993; Shannon, 1948]. Yet, particularly in the field of consciousness studies, the notion of information has acquired an increased autonomy, often on the verge of becoming something real that is acquired, stored, processed and transmitted [Aleksander and Gamez, 2010; Shanahan, 2010; Tononi, 2004]. This alleged autonomy of the information level has also received some support

in physics [Landauer, 1991]. Nevertheless, one can reasonably object that this way to address information and the mind is problematic. In fact, it is circular.

For historical reasons, humans have designed machines to perform computations [Gleick, 2011]. While such machines were originally rather simple, they have now achieved a level of complexity that exceeds our imagination. Yet, they are just an increasingly complicated version of the original blueprint: they are made of larger and larger networks of switches. Leaving aside the forthcoming quantum revolution, they are still nothing but networks of switches, ever faster and ever more. Calling them "computers" and attributing them the capacity to bring into existence "computations" is a questionable ontological step. We use them to compute and we gave to their switches an informational role. But are they really bringing into existence an additional level of reality? It seems highly questionable. In fact, we can use computations to describe any physical system (a star, a waterfall, a swarm of bees, a processor). The fact that we describe that physical system in computational terms does not change a physical system.

Suppose that there is a physical object, say a bunch of knotted ropes, and that such an object is suddenly used by a community of Inca accountants to carry out basic arithmetic operations, such as addition, subtraction, multiplication and division — as indeed happened with quipu. Would the bunch of ropes be any different because it is used by the Inca? Hardly. The ropes would just be ropes, no matter how they are used by us. Likewise, why should an electronic switch be any different because it is part of complicated networks that is used inside a cell phone by a human being?

Computationalism has been very popular among scientists and AI scholar because there is a long tradition associating thinking and computing — possibly as of Plato who suggested that mathematical ideas are the highest form of thought [Laraudogoitia, 2011]. In the last century, the same idea has been implicitly (and often explicitly) defended by mathematicians who gave birth to and later developed artificial intelligence [Arbib, 2003; Hopfield, 1982; Marr, 1982; Shannon, 1948; Taylor, 2007; Turing, 1950; Wiener, 1961]. More recently, many scientists [Chalmers, 2012; Davenport, 2012; Piccinini, 2016; Piccinini and Bahar, 2013]. Because of the ubiquitous success of computing devices, "The information processing metaphor of human intelligence now dominates human thinking, both on the street and in the sciences." [Epstein, 2016]. In this regard, [Piccinini, 2016, p. 203]:

> Computational explanation is so popular and entrenched that it's common for scientists and philosophers to assume computational theory of cognition without argument. But if we presuppose that neural processes are computations before investigating, we turn computational theory into dogma.

And a dogma it is in many cases. It is highly suspicious the fact that we have modeled the mental over the computational, mostly because in engineering department the main tool of enquire was a computer — when all you have is a hammer everything looks like a nail.

More recently, Giulio Tononi has even suggested that information, albeit in a special version christened integrated information, is a substance in itself to be added to the physical world [Aleksander and Gamez, 2010; Tononi, 2004; Tononi *et al.*, 1998]. While the notion of integrated information has gained some momentum and has led to interesting predictions [Boly *et al.*, 2011; Rosanova *et al.*, 2012], again there is no necessary connection between the physical substrate and the proposed conscious level. There is no necessity for integrated information to be conscious information. It might be so, but it would be a brute fact, something that has no explanation. More worryingly, the notion of integrated information, and by and large computationalism, is akin to a disguised form of dualism insofar as it suggests that there is the brain and there is consciousness. In the scientific community, computationalism's success seems to be due more to sociological factors than to actual evidence [Lau and Michel, 2019].

Yet this notion has been heavily criticized both by neuroscientists and AI experts [Brette, 2019; Epstein, 2016; Manzotti, 2012]. We don't know whether brain are information machines, we don't know if information is anything more than an abstract description of certain mechanisms, we don't know if information is anything that may have phenomenal qualities.

### 3.2. *Biological chauvinism*

Many authors have associated consciousness with life. It makes sense. As far as we know, so far, only living beings are conscious. But it is a necessary connection or is it only an historical accident? The main group of scholars defending this view is represented by that branch of enactivist that consider autopoiesis as a fundamental ingredient of the mind [Di Paolo, 2002; Froese and Taguchi, 2019; Froese and Ziemke, 2009; Thompson, 2007]. The connection is not obvious.

Of course, if biological chauvinism were true, machine consciousness would be doomed from the start. Fortunately, though, there is no evidence of any necessary connection between being alive and being conscious. After all there are many living organisms that are not conscious.

Enactivism needs to ground the notion of the agent on that of the living organism because otherwise it would be impossible to kickstart a series of key notions such as that of the body. Of course, it is not a real solution. It is only an epistemic promissory note. At least, enactivists are aware of this ontological principle. In this regard, Stewart states that "The paradigm of enaction solves this problem by grounding all cognition as an essential feature of living organisms [Stewart *et al.*, 2010]. By the same token, for Maturana and Varela, the great divide comes between matter and living organisms" [Maturana and Varela, 1980].

Of course, such a divide by suggesting an ontological difference between living organisms and other physical systems must be grounded. While many proponents of enactivism are supportive of such a difference [Di Paolo and Paolo, 2005; Di Paolo

*et al.*, 2017; Froese and Di Paolo, 2009; Thompson, 2004, 2007; Varela *et al.*, 1991], not everyone agrees. Recent works have tried to appeal to top-down causation or emergence [Aharonov *et al.*, 2018; Hoel *et al.*, 2013]. Yet these positions are all still very speculative. Why should a process be any different because it is part of a living organism rather than part of a mechanical system?

The same point we raised about the informational and the cognitive level holds for the living organism. It is a fact that the biological foundations of enactivism have had a great ontological cost that no has so far has been able to pay [Shani, 2020].

### 3.3. *Neural chauvinism*

Neural chauvinism is slightly different from biological chauvinism although the two views are obviously interconnected. The position suggests that biological neural networks have special emergent powers. John Searle endorsed this view when he defended the idea that the brain has "intrinsic intentionality" [Searle, 1992] but he has never explained why it should be so. Many scholars hanged on the intuitive notions that brains must be special or that biological neural networks must have special powers. Of course, if critically scrutinized, such a view amounts to little more than the infamous Muller's specific energies [Müller, 1826].

### 3.4. *Cognitivism*

Cognitive sciences have been a powerful paradigm to study and yet, when it comes to devise model for machines, its foundations aren't stronger than computationalism. Here too we have a level of description, namely cognition, and to work as an explanation of consciousness or as a foundation for machine consciousness, it must be taken as an emerging substance. By being taken as a substance, we refer here to something that must work as a physical cause or physical substrate of consciousness. Many scholars have defended cognition as a natural kind by means of conceptual gimmick [Adams and Aizawa, 2009; Aizawa and Adams, 2011], but the question remains.

So, as long as we look for the underpinning of consciousness, cognition is as epiphenomenal as it can be. In other words, it does not matter whether a machine is described using a cognitivistic terminology or not — the cogs and the micro circuits of the machine remain what they are: pieces of material stuff with no cognitive coating.

It is not by chance that cognition has often been associated with computation in the hope that, like Russell's tortoise, one may support the other [Arbib and Caplan, 1979; Chalmers, 2012; Piccinini, 2016]. Unfortunately, nobody has yet been able to provide an explanation as to why any of such descriptive levels should be causally relevant and physically real. If consciousness is a physical phenomenon it requires a physical underpinning. Cognition does not seem to have the necessary ontological support.

### 3.5. *Bodyism*

This is, unsurprisingly, a new word with a rather derogative meaning that one of the author has introduced elsewhere [Manzotti, 2019a]. It should address the dogmatic version of embodied cognition that looms behind most of the embodied cognition paradigm that has articulated in various schools (or versions) of embodied, embedded, extended and enacted cognition — Gallagher's 4E [Gallagher and Nelson, 2003; Newen *et al.*, 2018]. All such approaches suffer from the confusion between constitution and causation. While they have the resources to defend a causal role of embodiment and environment in shaping consciousness and cognition, they are far from showing convincingly that consciousness and cognition are constituted by the interaction between the body and the environment [Block, 2005].

For instance, while it is trivial to claim that the way in which we interact with the world has a role in shaping which properties are relevant in our experience, it is far from obvious that, say, my experience of opening a door is constituted by my movements. Moreover, constitution offers a problematic relation. More promising are recent attempts at supporting an identity between consciousness and sensori-motor interaction in the same spirit of the original enactivism [Beaton, 2016; Myin and Zahnoun, 2018; O'Regan and Noë, 2001].

Embodied cognition, as a foundation for the mind, suffers of several serious issues: circularity, epiphenomenalism, mentalism and disguised dualism [Manzotti, 2019a]. The principle problem is that they smuggle in the notion of body as an a priori explanatory principle. Unfortunately, this is not a feasible solution, because the body is not a primitive notion. In other words, a body is a body only insofar as it is owned by an agent.

This problem is particularly serious in the case of AI and robotics because it is immediately clear that attributing to our machines the status of having a body is misleading and arbitrary. In fact, does a washing machine have a body? Well, it is surely made of physical stuff, but is it enough to qualify as a body?

Likewise, the notion of the body, is circular relatively to other notions — e.g., agent, action, sensors, etc. If nobody were here, there would be no bodies. A rock does not have a body. Neither is a corpse a body. Bodies require agents. A body is a body only if there is somebody.

The intrinsic circularity of the notion of body jeopardizes the chance of success of em-body-ment as a theory of the mental. Embodiment entails body-ism that, in turn, entails some form of mentalism. Body-ism is the assumption that the body had a special status. For instance, calling "physical causes impinging on a body" stimulus is ontologically suspicious.

## 4. Current Approaches

It is fair to state that ever since the pioneering work of the founders of AI [McCarthy *et al.*, 1955; Shannon, 1948; Turing, 1950; Wiener, 1961], the goal of scientists has been to reproduce the human mind in its entirety. Human minds are conscious. Any

time that AI and robotics have shown a significant sign of progress, new hopes to address the original project has risen again.

After each wave of enthusiasm, the lack of concrete results induces many researchers to look with skepticism to the ones still struggling with existing methods [Harnad, 1990; Harnad and Scherzer, 2008; Lau and Michel, 2019; Searle, 1984]. Each wave of enthusiast researchers hoping to achieve consciousness in their generation is then followed by a corresponding wave of delusion and skepticism.

So far, the field of artificial consciousness has been around enough to be classified in various ways [Chella and Manzotti, 2009a; Gamez, 2008; Manzotti and Chella, 2018]. A first suggested distinction has been that between strong and weak machine consciousness mirrors that between strong and weak AI [Holland, 2003], and similar classifications have been proposed elsewhere.

Weak machine consciousness considers whether it is possible to build machines that behave as if they were conscious. Strong machine consciousness ventures to consider the possibility of actual conscious machines. Here, only the latter option is taken into consideration. Dodging the 'hard problem' is not a viable option in the business of making conscious machines.

A few scholars have suggested an intermediate approach, dubbed the "real problem of consciousness," namely addressing that its key might be "to recognize that explaining why consciousness exists at all is not necessary to make progress in revealing its material basis — to start building explanatory bridges from the subjective and phenomenal to the objective and measurable" [Seth, 2009]. According to Anil Seth, we should "account for the various properties of consciousness in terms of biological mechanisms; without pretending it doesn't exist (easy problem) and without worrying too much about explaining its existence in the first place (hard problem)." [Seth, 2016].

Yet, while the humility in this approach is undoubtedly reassuring, there are reasons to be skeptical. There are times in which prudence is not going to pay off — e.g., no precise measurement of the sun's position would have led to heliocentrism [Wollheim and Popper, 1935]. In the history of science, as we will mention in the end, whenever a truly intractable problem presented, scientists have not found the solution for free by minding their own business as usual. Big problems required significant changes in scientific premises.

The two approaches correspond loosely to Kuhn's incremental approach and extraordinary approach. According to the former (akin to Seth's real problem), achieving consciousness in machines is only a matter of incremental progress, as it is, says, playing chess. According to the latter, consciousness requires some conceptual breakthrough, that is, say, akin to Einstein's spacetime in contrast with Newton's traditional absolute space and time.

## 5. Future Goals and Requirements

In this paper, we won't outline our preferred direction. However, one of the authors has spent considerable time and resources trying to set aside a commonly accepted

idea, namely that consciousness is inside the head and that consciousness is not the outcome of computational processes [Manzotti, 2018, 2019a,b]. We won't recapitulate that hypothesis here. However, we want to outline what we expect the direction in which machine consciousness will likely develop might be

(1) Consciousness is a real phenomenon that apparently defies the scientific description of the physical world. The appearance of experience, even the more straightforward standard perception, is something totally unexpected.
(2) There are no reasons for desperation and conclude that consciousness cannot be understood and will forever be an unfathomable mystery [McGinn, 1999]. We are just at the beginning of this field, and thus it is likely that at this stage, the research is hampered by mistaken premises that, in Einstein's spirit, need to be revised.
(3) There is no reason to believe that consciousness is a phenomenon that requires new physical laws. It is something that takes place in our backyard, so to speak, caused by biological machinery. As far as we know, as complicated as the brain appears to be, it does not suggest any esoteric physical phenomena. The biological wetware is rather mundane in its underpinning physical phenomena; there are no apparent extreme temperature, speed, or physical phenomena.
(4) We must be careful not to fall in any form of homuncularism, geocentrism, anthropocentrism or dualism — namely position that solves the problem by attributing the brain of some extraordinary power, place or role. Consciousness is very likely much simpler and more common than is usually assumed.
(5) We must be careful from circular reasoning, as in the case of computationalism or vitalism. We must avoid question-begging approaches that assume what they should prove. Computationalism, vitalism, and embodied cognition are as many examples.
(6) There is no reason as to why consciousness might not occur as a result of artificially designed machines rather than natural-selection-derived biological organisms.
(7) The field of machine consciousness cannot confine itself only to technological proposals. Still, it must be brave enough to advance the theoretical proposition that will abide by the above points.

If such basic requirements are fulfilled, a solution will present itself. In particular, the field of machine consciousness cannot afford to be only a technological enterprise. Consciousness is something that runs afoul of the current scientific account of the physical world. Since this is not possible, consciousness is a familiar phenomenon occurring in a large number of animal species without particular efforts; it is clear that the clash between consciousness and physics must be only apparent. There must be something in our current model of physics that is too simplistic to account even for a mundane phenomenon such as our experience to occur.

In the history of science, this is not the first thing that something like that occurs. On the contrary, in the past, it has been the norm [Jammer, 1989; Kuhn, 1957, 1962; Reichenbach, 1958]. We have to remind that all scientific revolutions have been triggered by similar crisis: all of a sudden some phenomenon resisted any available explanations (a comet moving where and how it was not supposed to do so, the black body radiation, the photoelectric phenomenon, Mercury's perihelion precessions, the discrepancy between the alleged deep time of evolution and sun's apparent age, Michelson-Morley's finding of light speed constancy, the acceleration of the expansion of the universe, and so forth). Whenever a new phenomenon seems to challenge the existing conceptual framework, there are three possible approaches [Feyerabend and Maxwell, 1966; Kuhn, 1962]:

Pretending that the problem does not exist and continue to apply the existing methods in the hope that something will change (Einstein's definition of madness).

It is keeping the existing conceptual framework unchanged and adding additional hypotheses that address the problematic phenomenon without challenging the current conceptual order — e.g., the invention of the epicycles or, in the case of consciousness, suggesting the occurrence of additional emergent properties.

Revisiting the existing conceptual framework and suggesting a new order in which the problematic phenomenon will fit seamlessly without being a challenging anomaly — e.g., suggesting that all physical phenomena occur in discrete quantities [Einstein, 1905].

Consider for a moment the difference between the last two cases. When confronted with a phenomenon that does not fit in the existing conceptual order, the target community can react either suggesting new ad hoc hypotheses — e.g., integrated information, emergent properties, intrinsic intentionality — that do not really touch the foundation of the existing world view, or go back to the design room of science and revisit the roots of current theories — what did we miss?

So far, many of the approaches to machine consciousness have belonged either to the first or the second approach — a similar taxonomy has been advanced by [Gamez, 2008]. It is a fact that many scholars in AI, perhaps restrained by practical motivations [Lau and Michel, 2019; Michel *et al.*, 2019], have remained conservatively inside the established scientific picture. Possibly, also because AI is perceived as a technological endeavor that does not address the fundamental questions about the nature of the physical world but exploits them.

The fact is that artificial consciousness cannot be a purely technical discipline, as long as physics cannot be just a technical enterprise [Jammer, 1989]. There is a need for a new theoretical take on machine consciousness in the spirit expressed by Albert Einstein [Einstein, 1954]. Machine consciousness can achieve precisely that, by exploiting its hybrid nature of technological endeavor and theoretically challenging feat. So far, there has been a widespread consensus that only neuroscientists were allowed to do science and thus to make bold hypotheses about consciousness. There is no reason why machine consciousness should not do the same — gaining maturity

and autonomy as a scientific field by doing what all adolescents do, challenging the order they inherited from their parents, AI and Cognitive science.

## References


Adams, F. and Aizawa, K. [2009] *Why the Mind is Still in the Head* (Cambridge: Cambridge University Press).

Aharonov, Y., Cohen, E. and Tollaksen, J. [2018] Completely top−down hierarchical structure in quantum mechanics, *Proc. Natl. Acad. Sci.* **115**, 11730−11735, doi: 10.1073/pnas.1807554115.

Aizawa, K. and Adams, F. [2011] *The Bounds of Cognition* (Wiley, Singapore).

Aleksander, I. and Gamez, D. [2010] Informational theories of consciousness: A review and extension. in *BICS 2010*, Madrid.

Arbib, M. A. [2003] *The Handbook of Brain Theory and Neural Networks* (MIT Press, Cambridge).

Arbib, M. A. and Caplan, D. [1979] Neurolinguistic must be computational, *Behav. Brain Sci.* **2**, 449−483.

Beaton, M. [2016] Sensorimotor direct realism: How we enact our world, *Construct. Found.* **11**, 265−276.

Block, N. [2005] Review of Alva Noë's "Action in Perception", *J. Philos.* **102**, 259−272.

Boly, M., Garrido, M. I., Gosseries, O., Bruno, M.-A., Laureys, S., Friston, K. J., ... Litvak, V. [2011] Preserved feedforward but impaired top-down processes in the vegetative state, *Science* **332**, 858−862.

Brette, R. [2019] Is coding a relevant metaphor for the brain? *Behav. Brain Sci.* **42**(e215), 1−58.

Capurro, R. and Hjørland, B. [2005] The concept of information, *Annu. Rev. Inf. Sci. Technol.* **37**(1), 343−411, doi: 10.1002/aris.1440370109.

Chalmers, D. J. [1996] *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, New York).

Chalmers, D. J. [2012] A computational foundation for the study of cognition, *J. Cognit. Sci.* **12**, 1−20.

Chella, A. and Manzotti, R. [2009a] *Artificial Consciousness* (Thorverton).

Chella, A. and Manzotti, R. [2009b] Machine consciousness: A manifesto for robotics, *Int. J. Mach. Conscious.* **1**, 33−51.

Chella, A. and Manzotti, R. [2012] AGI and machine consciousness, *Theor. Found. Artif. Gen. Intell.* **4**, 263−282.

Davenport, D. [2012] Computationalism: Still the only game in town, *Minds Mach.* **22**, 183−190, doi: 10.1007/s11023-012-9271-5.

Di Paolo, E., Buhrmann, T. and Barandiaran, X. E. [2017] Di Paolo, E. A., Buhrmann, T. and Barandiaran, X. E. (eds.) *Sensorimotor Life: An Enactive Proposal* (Oxford University Press, New York).

Di Paolo, E. A. [2002] *Organismically-Inspired Robotics: Homeostatic Adaptation and Teleology Beyond the Closed Sensory-Motor Loop* (Adwance Knowledge International, Adelaide, Australia).

Di Paolo, E. A. and Paolo, E. [2005] Autopoiesis, adaptivity, teleology, agency, *Phenomenol. Cognit. Sci.* **4**, 429−452, doi: 10.1007/s11097-005-9002-y.

Dreyfus, H. L. [1972] *What Computers Can't Do: A Critique of Artificial Reason*, 1st ed. (Harper & Row, New York).

Dreyfus, H. L. [1992] *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press, Cambridge, MA).

Einstein, A. [1905] On the electrodynamics of moving bodies. *Ann. Phys.* **17**, 1−31.
Einstein, A. [1954] Foreword, in M. Jammer (ed.), *Concepts of Space* (Harvard University Press, Harvard).
Epstein, R. [2016] The empty brain. Your brain does not process information and it is not a computer. *Aeon*.
Feyerabend, P. K. and Maxwell, G. [1966] *Mind, Matter, and Method* (University of Minnesota Press, Minneapolis).
Froese, T. and Di Paolo, E. A. [2009] Sociality and the life−mind continuity thesis, *Phenomenol. Cognit. Sci.* **8**, 439−463, doi: 10.1007/s11097-009-9140-8.
Froese, T. and Taguchi, S. [2019] The problem of meaning in ai and robotics: still with us after all these years, *Philosophies* **4**(14), 1−14.
Froese, T. and Ziemke, T. [2009] Enactive artificial intelligence: Investigating the systemic organization of life and mind, *Artif. Intell.* **173**(3−4), 466−500, doi: 10.1016/j.artint.2008.12.001.
Gallagher, M. and Nelson, R. [2003] *Handbook of Psychology: Biological Psychology* (John Wiley & Sons, Hoboken, NJ).
Gamez, D. [2008] Progress in machine consciousness, *Conscious. Cognit.* **17**, 887−910.
Gleick, J. [2011] *The Information. A History, a Theory, a Flood* (Pantheon Books, New York).
Harari, Y. N. [2018] *21 Lessons for the 21st Century* (Spiegel & Grau, New York).
Harnad, S. [1990] The symbol grounding problem, *Phys. D* **1990**, 335−346.
Harnad, S. and Scherzer, P. [2008] First, scale up to the robotic turing test, then worry about feeling, *Artif. Intell. Med.* **44**(2), 83−89.
Hoel, E. P., Albantakis, L. and Tononi, G. [2013] Quantifying causal emergence shows that macro can beat micro, *Proc. Natl. Acad. Sci.* **110**, 19790−19795, doi: 10.1073/pnas.1314922110.
Holland, O. [2003] *Machine Consciousness* (Imprint Academic, New York).
Hopfield, J. J. [1982] Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. U. S. A* **79**, 2554−2558.
Huxley, T. H. [1863] *Man's Place in Nature* (Taylor & Francis, London).
Jammer, M. [1989] *The Conceptual Development of Quantum Mechanics* (American Institute of Physics, New York).
Kirk, R. [1974] Sentience and behaviour, *Mind* **83**(329), 43−60.
Kuhn, T. S. [1957] *The Copernican Revolution. Planetary Astronomy in the Development of Western Thought* (Harvard University Press, Harvard).
Kuhn, T. S. [1962] *The Structure of Scientific Revolutions* (The University of Chicago Press, Chicago).
Landauer, R. [1991] Information is physical, *Phys. Today* **44**, 23.
Laraudogoitia, J. P. [2011] Zeno and flow of information, *Synthese* **190**, 439−447, doi: 10.1007/s11229-011-0037-z.
Lau, H. and Michel, M. [2019] A socio-historical take on the meta-problem of consciousness, *PsyArXiv*, doi:10.31234/osf.io/ut8zq.
Manzotti, R. [2012] The computational stance is unfit for consciousness, *Int. J. Mach. Conscious.* **4**, 401−420.
Manzotti, R. [2018] *Consciousness and Object A Mind-Object Identity Physicalist Theory*, 1st ed. (John Benjamins Pub, Amsterdam).
Manzotti, R. [2019a] Embodied AI beyond embodied cognition and enactivism, *Philosophies* **4**, 1−15.
Manzotti, R. [2019b] Mind-object identity: A solution to the hard problem, *Front. Psychol.* **10**, 1−16, doi: 10.3389/fpsyg.2019.00063.

Manzotti, R. and Chella, A. [2018] Good old-fashioned artificial consciousness and the intermediate level fallacy, *Front. Robot. Artif. Intell.* **5**, 1−10, doi: 10.3389/frobt.2018.00039.
Marcus, G. [2014] The trouble with brain science, *Sci. Am.* 11.
Marr, D. [1982] *Vision* (Freeman, Francisco).
Maturana, H. R. and Varela, F. J. [1980] *Autopoiesis and Cognition: The Realization of the Living* (D. Reidel Pub. Co., Dordrecht, Holland).
McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. [1955] *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (Dartmouth College, Hanover, NH).
McGinn, C. [1999] *The Mysterious Flame. Conscious Minds in a Material World* (Basic Books, New York).
Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., ... Yoshida, M. [2019] Opportunities and challenges for a maturing science of consciousness, *Nat. Human Behav.* **3**, 104−107.
Michel, M., Fleming, S. M., Lau, H., Lee, A. L. F., Martinez-Conde, S., Passingham, R. E., ... Liu, K. [2018] An informal internet survey on the current state of consciousness science, *Front. Psychol.* **9**(2134), 1−5.
Müller, J. [1826] Vergleichende Physiologie des Gesichtssinnes.
Myin, E. and Zahnoun, F. [2018] Reincarnating the identity theory, *Front. Psychol.* **9**, 1−9, doi: 10.3389/fpsyg.2018.02044.
Newen, A., De Bruin, L. and Gallagher, J. S. (eds.) [2018] *The Oxford Handbook of 4E Cognition* (Oxford University Press, New York).
O'Regan, K. J. and Noë, A. [2001] A sensorimotor account of vision and visual consciousness, *Behav. Brain Sci.* **24**, 939−973.
Piccinini, G. [2016] The computational theory of cognition, in V. C. Muller (ed.), *Fund. Issues Artif. Intell.* (Springer, New York), pp. 203−221.
Piccinini, G. and Bahar, S. [2013] Neural computation and the computational theory of cognition, *Cognit. Sci.* **34**, 453−488.
Qvortrup, L. [1993] The controversy over the concept of information. An overview and a selected and annotated bibliography, *Cybern. Human Know.* **1**(4), 3−24.
Reichenbach, H. [1958] *The Philosophy of Space and Time* (Dover, New York).
Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M.-A., ... Massimini, M. [2012] Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients, *Brain* **135**, 1308−1320.
Schrödinger, E. [1958] *Mind and Matter* (University Press, Cambridge).
Searle, J. R. [1984] *Minds, Brains, and Science* (Harvard University Press, Cambridge, MA).
Searle, J. R. [1992] *The Rediscovery of the Mind* (MIT Press, Cambridge, MA).
Seth, A. K. [2009] The strength of weak artificial consciousness, *Int. J. Mach. Conscious.* **1**, 71−82.
Seth, A. K. [2016] The real problem, *Aeon* **11**, 1−11.
Shanahan, M. P. [2010] *Embodiment and the Inner Life. Cognition and Consciousness in the Space of Possible Minds* (Oxford University Press, Oxford).
Shani, I. [2020] Befuddling the mind: Radical Enactivism (Hutto-Myin style) and the metaphysics of experience, *Phenomenol. Cognit. Sci.* 1−18.
Shannon, C. E. [1948] A mathematical theory of communication, *Bell Syst. Tech. J.* **27**, 379−423, 623−656.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... Hassabis, D. [2017] Mastering the game of Go without human knowledge, *Nature* **550**, 354−359.
Stewart, J., Gapenne, O., Di Paolo, E. A. and Paolo, E. A. D. [2010] *Enaction* (The MIT Press, Cambridge, MA).

Taylor, J. G. [2007] CODAM: A neural network model of consciousness, *Neural Netw.* **20**, 983−992, doi: 10.1016/j.neunet.2007.09.005.

Thompson, E. [2004] Life and mind: From autopoiesis to neurophenomenology. A tribute to Francisco Varela, *Phenomenol. Cognit. Sci.* **3**, 381−398.

Thompson, E. [2007] *Mind in Life. Biology, Phenomenology, and the Sciences of Mind* (The Belknap Press of the Harvard University Press, Cambridge, MA).

Tononi, G. [2004] An information integration theory of consciousness, *BMC Neurosci.* **5**, 1−22.

Tononi, G., Edelman, G. M. and Sporns, O. [1998] Complexity and coherency: Integrating information in the brain, *Trends Cognit. Sci.* **2**, 474−484.

Turing, A. M. [1950] Computing machinery and intelligence, *Mind* **59**, 433−460.

Varela, F. J., Thompson, E. and Rosh, E. [1991] *The Embodied Mind: Cognitive Science and Human Experience* (MIT Press, Cambridge, MA).

Wiener, N. [1961] *Cybernetic: Or Control and Communication in the Animal and the Machine* (John Wiley & Sons, Cambridge, MA).

Wollheim, R. and Popper, K. R. [1935] *The Logic of Scientific Discovery* (Verlag von Julius Springer, Vienna).

Yasnitsky, L. N. [2020] Whether be new "Winter" of artificial intelligence? in *Paper Presented at the Integrated Science in Digital Age. ICIS 2019*, Dordrecht.