International Journal of Machine Consciousness
Vol. 1, No. 1 (2009) 33–51
© World Scientific Publishing Company

MACHINE CONSCIOUSNESS: A MANIFESTO FOR ROBOTICS

ANTONIO CHELLA

Department of Computer Engineering,
University of Palermo, Viale delle Scienze, Palermo, 90128, Italy
chella@unipa.it

RICCARDO MANZOTTI

Institute of Consumption, Communication and Behavior,
IULM, Via Carlo Bo 1, Milan, 20143, Italy
riccardo.manzotti@iulm.it

Machine consciousness is not only a technological challenge, but a new way to approach scientific and theoretical issues which have not yet received a satisfactory solution from AI and robotics. We outline the foundations and the objectives of machine consciousness from the standpoint of building a conscious robot.

Keywords: Robot consciousness; embodiment; situatedness; representation; externalism; robotics.

1. Machine Consciousness: A Cross-Road Among Disciplines Imposing Old and New Questions

The recent upsurge of interest for modeling and implementing conscious machines or conscious robots[1–10] is motivated by the belief that machine consciousness will shed new light on the many critical aspects lurking in AI and robotics.

Machine consciousness is not simply a technological challenge, but a field posing theoretical and scientific issues such as the relations between information and meaning, the nature of teleology, the unity of the self, the integration of information, the nature of experience, and many others from a novel point of view.

Machine consciousness has a long past and a very brief history.[11] Although the term is fairly recent,[12] the problem has been addressed since Leibniz's mill at least. Machine consciousness offers the opportunity to deal with the hard problem of consciousness from a different perspective — a fact already clear 40 years ago when Hilary Putnam wrote that:

> "*What I hope to persuade you is that the problem of the Minds of Machines will prove, at least for a while, to afford an exciting new way to approach*

> *quite traditional issues in the philosophy of mind. Whether, and under what conditions, a robot could be conscious is a question that cannot be discussed without at once impinging on the topics that have been treated under the headings Mind-Body Problem and Problem of Other Minds.*"[13]

Machine consciousness is an outrageous field of enquiry for at least two reasons. First, it takes consciousness as a real phenomenon with practical effects on behavior, a standpoint that has raised eyebrows until a few years ago but is now accepted in many scientific areas.[14–17] Secondly, it suggests the possibility to reproduce, by means of a man-made machine, the most intimate of mental aspects — namely conscious experience. Although many argued against the possibility of machine consciousness on the basis of either *a priori* or ideological reasons ("no machine will ever be like a man"), so far no one has conclusively argued against such a possibility. Biological chauvinism is not at all based on sound arguments.

Besides, arguments which deny the possibility of machine consciousness typically would deny the very possibility of human consciousness whether applied to human beings. Let us expand this point. A naïve adversary could argue that a robot could never be conscious, since CPUs and computer memory do not seem to be the right kind of stuff to harbor phenomenal experience. And yet, borrowing Lycan's words, if such

> "*...pejorative intuition were sound, an exactly similar intuition would impugn brain matter in just the same way* [...]: '*A neuron is just a simple little piece of insensate stuff that does nothing but let electrical current pass through it from one point in space to another*; *by merely stuffing an empty brainpan with neurons, you couldn't produce qualia-immediate phenomenal feels!*' — *But I could and would produce feels, if I knew how to string the neurons together in the right way; the intuition expressed here, despite its evoking a perfectly appropriate sense of the eeriness of the mental, is just wrong.*"[18]

Yet, where does consciousness stem from? How can a brain be conscious? Contrary to AI and functionalism, some scholars regard the functional view of the mind insufficient to endorse the design of a conscious robot. Here, the commonly raised arguments against strong AI, based on the reduction of machines to Turing machines, lose some of their strength — Searle's Chinese Room and Block's Chinese nation arguments being the two most famous cases. In fact, a machine is not necessarily a Turing machine. Although most available machines are instantiation of von Neumann's blue print, other architectures are becoming available and more are going to be designed and implemented in the near future. There is no reason why all machines should be instantiations of Turing machines, as it has been observed, "One day, perhaps sooner than people think, [artificial agents] development may take place in Petri dishes or quantum computers, not CPUs."[19]

Furthermore, it is an open issue whether robots are reducible to a purely mechanistic view, when considered as part their environment. Standpoints such as embodiment, situatedness, and integration challenge the classic AI disembodied view of a syntactical symbol-crunching machine,[20–23] as pointed out by Dretske:

> "*Work on machine perception, pattern recognition, and robotics has greater relevance to the cognitive capacities of machines than the most sophisticated programming in such purely intellectual tasks as language translation, theorem proving, or game playing.*"[24]

Roughly speaking, machines consciousness lies in the promising middle ground between the extremes of biological chauvinism (i.e., only brains are conscious) and liberal functionalism (i.e., any behaviorally equivalent functional systems is conscious). Machine consciousness proponents maintain that biological chauvinism is too narrow and yet they concede that some kind of physical constraints will be unavoidable (hence no multiple feasibility) in order to build a conscious agent.

Recently, some authors emphasized the alleged behavioral role of consciousness[20,25–27] to avoid having the problem of phenomenal experience. Holland suggested that it is possible to distinguish between *weak artificial consciousness* and *strong artificial consciousness.*[9] The former approach deals with agents which behaved *as if* they were conscious, at least in some respects. Such a view avoids any commitment to the hard problem of consciousness. Instead, the latter approach deals squarely with the possibility of designing and implementing agents capable of real conscious feelings.

Although the distinction between weak and strong artificial consciousness sets a temporary working ground, it suggests a misleading view. Setting aside the crucial feature of the human mind — namely experience, i.e., phenomenal consciousness — misses something indispensable for the understanding of cognition. Skipping the so called "hard problem" is not a viable option in the business of making conscious machines.

Further, the distinction between weak and strong artificial consciousness is misleading because it mirrors a dichotomy between true conscious agents and "as if" conscious agents. Yet, human beings are conscious and there is evidence that most animals exhibiting behavioral signs of consciousness are phenomenally conscious. It is a fact that human beings have phenomenal consciousness. They experience pains, pleasures, colors, shapes, sounds, and many more other phenomena. They feel emotions, feelings of various sort, bodily and visceral sensations. Arguably, they also have phenomenal experiences of thoughts and of some cognitive processes.[28,29] In sum, it would be very bizarre whether natural selection had gone at such great length to provide us with consciousness if there was a way to get all the advantages of a conscious being without actually producing it. Thus we cannot help but wonder whether it is possible to design a conscious machine without dealing squarely with the hard problem of consciousness. Hence, the dichotomy between weak and strong artificial consciousness is highly probably fictitious.

While some authors adopted an open approach that does not rule out the possibility of actual phenomenal states in current or future artificial robots,[1,8] other authors[30,31] maintained that a conscious machine is necessarily a phenomenally conscious machine. For them, to be conscious is necessarily having phenomenal experiences.[32] For instance, Tononi suggested that the kind of information integration necessary to exhibit the kind of behavioral unity and autonomy of a conscious being is also associated to certain intrinsic causal and computational properties which is responsible for phenomenal experience.[33] It is still an open issue whether having phenomenal consciousness is a requisite or an effect of a unique kind of cognitive architecture.

2. Theoretical and Scientific Issue at the Roots of Machine Consciousness

In this section, we sketch an impressionistic overview of the scientific, theoretical and philosophical issues at the roots of machine consciousness (indeed often of consciousness itself). Too often, researchers in the field accept assumptions which are far from being justified either empirically or theoretically. As a result, many years are wasted in pursuing goals on the basis of unwarranted premises. We mention two chasms which are often presented as definitive obstacles to the possibility of a conscious machine.

The first chasm is the separation between artificial and natural entities. Such a separation further unfolds in a various ways: either between man-made, and no-man-made, or between inorganic and biological. With regard to the conscious mind, they are both hugely overestimated. First, no one has ever been able to suggest any kind of necessary link between the carbon-based molecules featured by living organisms and consciousness. At a meeting sponsored in 2001 at the Cold Spring Harbour Laboratories addressing the question "Could Machines Be Conscious?", the participants agreed on the fact that there is no known law of nature that forbids the existence of subjective feelings in artifacts designed or evolved by humans. On the other hand, living human beings belong to the physical domain and it is fair to consider them conscious. Hence, the physical world can host consciousness. A machine could exploit the same mechanism. So much for the first chasm.

The chasm between mental and physical domains is a lot harder. Luckily, machine consciousness is, once again, extremely useful for two reasons. First, it compels to reconsider critically several notions which are used rather naively in neuroscience and cognitive science such as representation, information, meaning, intentionality, mental images and so on. Secondly, it requires either dropping all form of theoretically unsupported forms of "bio prejudices" or somehow justifying them. In other words, studying consciousness in humans may be misleading since it is easy to slip in the scientifically unsupported conviction that humans are somehow special and hence implicitly assuming some hidden power.

We believe it is useful to consider at a glance the following list of issues for various reasons. The list shows how complex the issue of consciousness is. Secondly, many of these issues are deep scientific and theoretical problems and not simply technical challenges. Finally, such a sketchy overview will make plain that most (if not all) of these problems are deeply interrelated together. The emerging general picture is that of a yet-to-be-defined framework that will probably take advantage of a needed theoretical twist.[34]

2.1. *Embodiment*

A much heralded crucial aspect of agency is embodied cognition.[35–38] Although, it cannot be underestimated the importance of the interface between a robot and its environment, as well as the importance of an efficient body, it is far from clear whether this aspect is intrinsically necessary to the occurrence of consciousness. Embodiment does not seem to be a sufficient condition for consciousness. Arguably, embodiment could be a necessary condition.

We do agree that a body is indeed a necessary condition, yet we wonder whether there had been any clear understanding of what embodiment is. Apart from intuitive cases, when is an agent truly embedded? On one hand, there is no such a thing as a "unembodied agent", since even the most classic AI algorithm has to be implemented as a physical set of instructions running inside a physical device. On the other hand, even a complex robot such as ASIMO is not really embodied. It has a centralized inner controller whose behavioral rules are hard-wired by its designers.

There are many biological agents that would apparently score very well on embodiment but yet do not seem good candidate for consciousness. Take insects, for instance. They show impressive morphological structures that allow them to perform outstandingly well without very sophisticated cognitive capabilities.[37,39]

The notion of embodiment is a lot more complex than the simple idea of having a body and controlling actuators and sensors. It refers to the kind of development and causal processes engaged between a robot, its body, and its environment.[38,40] So far, no thorough analysis has been presented.

2.2. *Situatedness*

Besides having a body, a conscious agent needs also to be part of a real environment, i.e., being situated. Yet the necessity of situatedness is not totally uncontroversial. For instance, many authors argued that consciousness is a purely virtual inner world created inside a system which, to all respects, lacks any direct contact with the environment.[41–43] They advocate the actual possibility of a conscious brain in a vat. Yet we do not have any empirical evidence that an "unsituated brain" would ever be conscious. There are no actual examples to quote it. To this extent the possibility of a pure virtual phenomenal experience is bizarre, and this bizarreness dims its appeal considerably.

If consciousness requires embodiment and situatedness, we should be able to point out — and we are *not*, at the present time — what is to be situated. What kind of architecture and individual history is sufficient for situatedness?

Usually, alleged embodied robots such as Paul or Brook's agents, Babybot, passive walkers, and similar robots[44–47] are regarded as examples of integration with the environment since they outsource part of their cognitive processes to smart morphological arrangements that allow greater efficiency or simpler control.

Yet this is controversial. Situatedness involves some kind of developmental integration with the environment such that what the robot is and does is a result of the past interactions with the environment. A real integrated agent is an agent that changes in some non trivial way (which has to be better understood) as a result of its tight coupling with the environment. The aforementioned artificial agents lack this kind of development: they remain more or less the same notwithstanding their interplay with their environment. Their teleological structure is unchanged by their interactions.

Another fruitful approach is represented by those implementations that outsource part of the cognitive processes to the environment[48] and explicitly consider the agent as a part of the environment. For instance, the field of epigenetic robotics is strongly interested in designing robots capable of developing accordingly with the environment.[49–51]

2.3. *Emotions and motivations*

It has been maintained that emotions could be the key to consciousness. Damasio suggested that there is a core consciousness supporting the higher forms of cognition.[52] Although this is a fascinating hypothesis, it remains unclear how emotions should be implemented. Although many roboticists draw inspiration from various emotional models,[53–60] whether an architecture could be considered "emotional" is very far from being clear.

Further, it is possible that consciousness is prior to emotions. So far, what has been called "emotions" in robots is more akin to a smart cognitive shortcut. Instead of learning how to behave on the basis of a general algorithm, researchers injected bundles of heuristic thumb-rules disguised as emotional modules into their robots. The result is behaviorally effective but it is not necessarily a real progress in understanding what emotions are. In robotics they are often not more than wishful labeling of modules (see Sec. 3.2).

For instance, a confusing approach is offered by certain descriptions of Kismet as a robot with emotions.[58] Kismet has nothing to do with emotions apart from mimicking them in front of their users. The robot does not contain any convincing model of emotions but only an efficacious hard-wired set of rules to control its captivating robotic human-like facial features. In Kismet case, it is not altogether wrong saying that emotions are in the eye of the human beholder.

2.4. *Unity and integration*

Consciousness seems to be deeply related with the notion of unity. Yet what gives it unity to a collection of parts, being them events, parts, processes, computations, instructions? The ontological analysis has not gone very far[61,62] and the neuroscientists wonder at the mystery of neural integration often labeled as the binding problem.[63,64] Machine consciousness has to face the same issue. Would it be enough to provide a robot with a series of capabilities for the emergence of a unified agent? Should we consider explicitly the necessity of a central processing locus or would unity stem out of some other completely unexpected aspect?

Classic theories of consciousness are often vague as to what gives unity. For instance, would the *Pandemonium*-like community of software demons championed by Dennett[65] gain a unity eventually? Has a simple software gained unity apart from the programmer's intentions? Would embodiment and situatedness be helpful?

A possible and novel approach to this problem is the notion of integrated information introduced by Tononi.[33] According to him, certain ways of processing information are intrinsically integrated because they are going to be implemented in such a way that the corresponding causal processes get entangled together. Although still in its initial stage, Tononi's approach could cast a new light on the notion of unity in a cognitive agent.

2.5. *Time*

Conscious experience is located in time. We experience the flow of time in a characteristic way which is both continuous and discrete. On one hand, there is the flow of time in which we float seamlessly. On the other hand, our cognitive processes require time to produce conscious experience and they are located in time. Surprisingly, there is evidence showing that we are visually aware of something only half a second after our eyes have received the relevant information.[66]

The classic Newtonian time fits very loosely with our experience of time. According to Newton, only the instantaneous present is real. Everything has to fit in such Euclidean temporal point. For instance, speed is nothing more than the value of a derivative and can be defined, pace Zeno, at every instant. We are expected to occupy only an ever-shifting temporal point with no width. The Einstein−Minkowsky space-time model is not particularly enlightening in this respect.[67] Time remains a dimension in which the present is a point with no width. Such an instantaneous present cannot accommodate the long lasting and content-rich conscious experience of present.

Neuroscience faces similar problems. According to the neural view of the mind, every cognitive and conscious process is instantiated by patterns of neural activity. This apparently innocuous hypothesis hides a problem. If a neural activity is distributed in time (as it has to be since neural activity consists in temporally distributed series of spikes), there must be some strong sense in which something taking place in different instants of time belong to the same cognitive or conscious process. For instance, what glues together the first and the last spike of a neural activity leading a subject to perceive a face? Simply suggesting that they occur inside the same window of neural activity is like explaining a mystery with another mystery. What is a temporal window? And how does it fit with our physical picture of time? Indeed, it seems to be at odds with the instantaneous present of physics.

In the case of robots, this issue is extremely counterintuitive. For instance, let us suppose that a certain computation is identical with a given conscious experience. What would happen if we purposefully slow down the taking place of the same computation? Certainly, we can envisage an artificial environment where the same computation is performed at an altered time (for instance we could simply slow down the internal clock of such a machine). Would the alleged conscious robot have identical conscious experience but spread in a longer span of time?

A related issue is the problem of the *present*. As in the case of brains, what defines a temporal window? Why are certain states part of the present? Does it depend on certain causal connections with behavior or is it the effect of some intrinsic property of computations?

We have no answer to such questions, but this is not a good reason not to ask them. Further, it is even possible that we would need to change our basic notion of time.

2.6. *Free will*

Another classic issue which does not fit with our mechanistic picture of a machine is the fact that a conscious robot ought to be capable of a unified will often assumed as *free*. The topic is as huge as a topic can be (for a comprehensive review see Kane[68]).

A classic argument against free will in human and hence, *fortiori*, in a machine is the following.[69] If a subject is nothing but the micro-particles constituting it (and their state), all causal powers are drained by the smallest constituents. In other words, you and I cannot have a will different from what all the particles constituting us do.[70] If the argument holds, there will be no space left for any high level causal will. All reality ought to reduce causally to what is done by the micro-particles who would be in total charge of what happens. No top-down causation would be possible and no space would remain for the will of a subject to interfere on the course of events.

Yet, we have a strong intuition (perhaps wrong) that we are capable of willing something and that our conscious will is going to make a difference in the course of events. Many philosophers strongly argued in favor of the efficacy of conscious will. If such view would run afoul against our theories of mental causation, so much the worse for them.[71]

Another threat to free will allegedly comes from Libet's famous studies, in which it was shown that we are conscious of our choices only after 300 ms our brain has made them.[72] Although Libet left open the possibility that our consciousness can veto the deliberations of our brains, there is still a lot of controversy about the best interpretation of his experimental results.

In short, a huge open problem is whether a system *as a whole* can have any kind of causal power over its constituents. Since consciousness seems to be strongly related with the system as a whole, we need some theory capable of addressing the relation between wholes and parts.

As to robots, the issue is difficult as ever. The classic mechanistic approach and several respected design strategies (from the traditional *divide et impera* rule of thumb, to sophisticated object-oriented programming languages nowadays) suggest to conceive machines as made of separate and autonomous modules. Then, robots would be classic examples of physical systems where the parts completely drain the causal power of the system as a whole. From this point of view, robots would be completely unsuited to endorse a conscious will. However, there are some possible approaches that can provide a viable route out of this blind alley.

An approach is based on recent connectionist proposals that stressed the kind of connectivity between elementary computational units. According to such proposals it is possible to implement networks whose behavior is not reducible to any part of the network, but rather it stems out of the integrated information of the system as a whole.[33]

Another approach stresses the teleological roles of certain feedback loops that could do more than classic control feedbacks. Here, the idea is to implement machines capable of modifying their teleological structure in such a way as to pursue new goals by means of a tight coupling with their environment. Thus, the behavior of a robot would be the result of all its past history as a whole. There would not be separate modules dictating what the robot has to do, but rather the past history as a whole would reflect in every choice.[40]

In short, a robot does not need to be so mechanistic, after all. Once more, machine consciousness could help us to further push the limits of our understanding of classic notions.

2.7. *Representation*

Representation is one of the most controversial problems. How is it possible that something represents something else? We face an apparent insurmountable problem: the physical world is defined in term of extensional entities which do not refer to anything.

As a result the physical world cannot possess any semantics. In fact, semantics has been charged to conscious subjects without suggesting any convincing explanation as to how subjects could emerge out of a physical world. The classic Searle's argument suggests that robots do not have intrinsic intentionality and thus are devoid of semantics. If this was true, robots would not ever be conscious since they are syntactic engines. However this view does not explain why brains are special with regard to representation (and intentionality). Searle's suggestion that brains have special causal powers has never been too persuasive.

Hence, it is possible that we need to reframe our view about the physical world in order to accommodate the apparently impossible fact of representation. All attempts at naturalizing semantics, intentionality, and representation (with all the well-known differences among these terms) either failed or did not succeed enough.[73–76] How can symbols be grounded with other facts in the world?[77,78] Yet, it is a fact that our minds are capable of representing the external world.

It is curious that neuroscience is tempted by the metaphors introduced by computer science in order to provide (incomplete) explanations of the activity of the brain.[79] The current debate about the existence of a neural code or about mental imagery are deeply indebted with the computer science view of the mind. Why should there be a code in the brain and why should a code provide any justification of the brain semantics if no code in machines seems to be able to pay the bill? In short, any argument that applies different criteria to biological and artificial agents should be rejected.

2.8. *Experience*

Finally, the problem which is apparently the most conspicuous, is how can a physical system produce anything like our subjective experience, i.e., qualitative phenomenal content? At sunset, our retinas are hit by light rays and we experience a glorious symphony of colors. We swallow molecules of various kinds and, as a result, we feel the flavor of a delightful *Brunello di Montalcino* red wine. Or so, it is usually thought. Harnad and Scherzer wrote that:

> "*Consciousness is feeling, and the problem of consciousness is the problem of explaining how and why some of the functions underlying some of our performance capacities are felt rather than just 'functed'.*"[80]

Galileo Galilei famously suggested that smells, tastes, colors, and sounds are nothing outside the body of a conscious subject.[81] Experience is allegedly created by the subject body in some unknown way. A very deep rooted assumption is the separation between the domain of experience, i.e., subjective phenomenal content, and the domain of objective physical events. Such assumption is intertwined with the epistemological roots of science itself. It is a dogma the claim that physical reality can be adequately described from a quantitative third-person perspective oblivious of any qualitative aspect. After all, if you read a handbook of physics, you will find mainly mathematical equations describing a purely quantitative view of reality. There is no space left for qualitative phenomenal feelings.

Yet many scientists and philosophers alike questioned the soundness of such a distinction as well as our true understanding of the nature of the physical.[82–87]

Whether the mental world is a special construct concocted by some irreproducible feature of the nervous systems of most mammals, is still an open question. It is fair to stress that there is neither empirical evidence nor theoretical arguments supporting such a view. In the lack of a better theory, we take into consideration the rather surprising idea that the physical world comprehends also those features that we usually attribute to the mental domain.[84,88] A honest physicalist must be held that if something is real, and we assume consciousness is real, it has to be physical. Hence, in principle, a device can envisage it.

In the case of robots, how is it possible to take over the "functioning versus feeling" divide?[18,80] As far as we know, a robot is nothing more than a collection of interconnected modules, each functioning in a certain way. Why should the functional activity transfigure in the feeling of a conscious experience? However, the same question could be asked about the activity of neurons. Each neuron, taken by itself, does not score a lot better than a software module or a silicon chip as to the emergence of feelings. At least, as far as we know. So one possibility remains: It is not a problem of the physical world but rather of our picture of the physical world. Couldn't we discount a too simplistic view of the physical world? Robots are part of the same physical world which produced conscious human subjects, thus they could take advantage of the same relevant properties and features.

2.9. *Other issues*

We admit that there is a very long list of correlated issues, that are not adequately addressed here: intentionality, 1st person versus 3rd person perspectives, qualia, relation between phenomenal content and knowledge, exotic physical phenomena described by quantum laws, mental imagery, meaning, symbol grounding, and so on. Although some of them overlap partially with the topics in the previous sections, some others are peculiar in their own. However, most issues share a similar structure with regard to machine consciousness — as long as some arguments seem to prevent a robot from being conscious, the same argument would prevent a brain to be so. Yet, human beings are conscious and thus there must be some mistake in our assumptions about that argument. If an argument against machine consciousness should reject consciousness in humans too, so much the worse for the argument.

For instance, we can deny that robots will ever be conscious since they are made of physical stuff which is alleged to be devoid of phenomenal content. Yet, brains are made of physical stuff, too. On the same basis, are we really going to deny that human beings are conscious? Hardly. Rather we should question those assumptions that lead us to such a conclusion.

3. Common Mistakes and Open Issues

We left for the end of the paper a quick analysis of critical aspects. We would like to point out a series of possible common mistakes that could affect many otherwise fruitful approaches. We tried to focus only on those possible methodological mistakes that are specific to the field of machine consciousness in order to build a conscious robot.

3.1. *False goals*

Due to its vagueness and intrinsic difficulty, consciousness has often been reduced to a more tractable aspect. This is an example of the "mereological fallacy" in which a problem is identified only with a part of itself.[79] For instance, it is true that a conscious agent is often also an autonomous agent. However, are we sure that an autonomous agent is necessarily a conscious one? Similar arguments suggest a more cautious approach for other capacities and aspects presented as sufficient for conscious experience: autonomy, embodiment, situatedness, resilience, and so on.

Whether or not consciousness can be reduced to certain capacities or features that are often correlated with the existence of a conscious agent is, to say the least, obscure. Along these lines, Tononi and Koch maintained that consciousness does not require many of the skill that many scholars are trying to emulate in machines:

> "*Remarkably, consciousness does not seem to require many of the things we associate most deeply with being human*: *emotions*, *memory*, *self-reflection*, *language*, *sensing the world*, *and acting in it.*"[31]

The issue remains controversial. Most scholars would probably argue against such a view — more prominently those that associate conscious agency with the capacity either to integrate cognitive skills[20,25,89] or to be autonomous, resilient, embodied,[26,49] and situated.[87,90]

3.2. *Labeling*

Very often cognitive scientists, roboticists and AI researchers present their architecture, labeling their boxes with intriguing and suggestive names: "emotional module", "memory", "pain center", "neural network", and so on. Unfortunately, labels on boxes in diagrams constitute empirical and theoretical claims that have to be justified elsewhere. To use Dennett's terminology they are "explanatory debts that have yet to be discharged".[91]

Even a rather uncontroversial term such as "neural network" is loaded with vague references to troublesome issues. The very choice of the name endorsed a series of expectations. If we had not been under the "spell of some psychologist's intuition" and if we had known the same computational tool under a more sober name such as "not linear functional approximator", the explanatory expectations would have been a lot less appealing.

A frequent, often reasonable, complaint from machine consciousness skeptics addresses the liberal use of demanding and not always fully justified labels.

3.3. *Levels*

It is easy to accept the existence of multiple level of reality co-existent in the same physical system. Why should we not talk of bits or numbers (or even images and sounds) when referring to the content of computer memories? However, it is well known that such use could be a powerful source of confusion.[79]

For instance, are there images in a computer memory? From a physical perspective, there are different levels of tensions in small capacitors. From another perspective, there are logical values in logical gates. Getting higher and higher, we obtain bits, numbers, array, RGB triplets, and images. We could get even more abstract and consider the existence of images having a certain content. But are all these levels real or are they just different perspectives on the same phenomenon? With regard to this issue, the trouble is that most of these levels — bits, logical values, numbers, RGB triplets — are properties of a way of thinking about what takes place in our computer; they are not properties of the computer. What we take to be two *pixels* in an image is nothing but two tensions causally related with what happens on a computer screen. Pylyshyn wrote:

> "*The point here is not that a matrix representation is wrong. It is just that it is neutral with respect to the question of whether it models intrinsic* (*i.e., architectural*) *properties of mental images.*"[92]

In the case of machine consciousness, the problem cannot be postponed since there is, at least, one level that should be real: the level of conscious experience. But why?

3.4. *Metaphors*

On a similar note, we ought to be suspicious of many metaphors that populate our technical and scientific jargon such as "computation", "information", "system", "symbol", and "state of a system". Although they are very useful shortcuts whose value cannot be underestimated, do they correspond to real physical phenomena?

For instance, take the notion of the state of a relatively simple system such as a planet. It could correspond to a vector measuring various magnitudes: position, speed, moment, acceleration, and so on. There are at least two curves now: the physical one of the planet in space and the curve of the state of the system in its multidimensional space state. Yet only the former is physical and actual while the latter is a sophisticated way to express the former. The space in which the state is moving is not real. It is an abstraction.

This is one example of what Alfred Whitehead defined as the fallacy of the *misplaced concreteness*. We mistake a conceptual abstraction for a concrete entity:

> "[*The Fallacy*] *is merely the accidental error of mistaking the abstract for the concrete... This fallacy is the occasion of great confusion.*"[93]

Once more, machine consciousness could offer an opportunity to clarify such issue.

3.5. *Simulation*

At some point in the development of conscious theories, scholars should tackle the issue of the relation between simulation and simulated. Although everybody does agree that simulated waterfalls are not wet, intuition fails as to whether a simulated conscious feeling is felt.

Although it has seldom been explicitly discussed, many hold that if we could simulate a conscious agent, we would have conscious agents. Is a simulated pain painful?

From a physical perspective, a simulation is a physical system that bears some analogy with another physical system. For instance, the simulation of a waterfall produces on my screen a series of pixels which moves analogously to the droplets of a real waterfall. More poignantly, the analogy could refer to more subtle functional, relational, and structural properties. But the analogy is an epistemic choice like all relations of analogy or similarity. Neither similarity nor analogy are adequate criteria to endorse implementation.

3.6. *Inside and outside: internalism versus externalism*

There are two powerful conceptual attractors in the discussion on consciousness. They are going to exert their strength in the machine consciousness arena, too. Where is the mind and its content located? Inside or outside the body of the agent? So far, neither option proved entirely satisfactory and the debates keeps running.

On one hand, it would be very simple if we could locate consciousness inside the body of the agent and thus inside our future conscious robots. For instance, according to Kim, "if you are a physicalist of any stripe, as must of us are, you would likely believe in the local supervenience of qualia."[94] That the mind must somehow depend on what take place *inside* the body. However, such a view is not convincing since mental states (broadly speaking) are about something that often appears as being internal to the body.[87,90,95–98] Therefore, mental states should somehow address external states of affairs (whether they are concepts, thoughts, percepts, objects, events). Unfortunately, there are no available theories explaining how the arrow of the mind could hit the external world and, consequently, many authors opted for a completely internal view of the mind. Since the world cannot get in, the mental world must be inside the agent from the beginning or it must be concocted inside.[41,99] All these positions can broadly be labeled as cases of *internalism*.

On the other hand, consciousness refers to the external world. Maybe, it is so difficult to bring content inside the mind because it remains outside. So we should reframe our model of the agent such as to include the external world.[87,95,97] Such a twist in our perspective about the limit of the robot would endorse those views that consider embodiment and situatedness as necessary conditions for a conscious robot.

4. Conclusion

In this paper we discussed the issues to be addressed in order to design and build a conscious robot such as embodiment, situatedness, emotions and motivations, unity, time, free will, representation, and qualitative experience. We also discussed some of the common mistakes and pitfalls.

Mainstream Artificial Intelligence addressed these issues only slightly. Although AI achieved impressive results,[100] it is always astonishing the degree of overestimation that many non experts seem to stick to. In 1985, while addressing the American Philosophical Association, Fred Drestke was sure that "even the simple robots designed for home amusement talk, see, remember and learn".[24] It is not unusual to hear that robots are capable of feeling emotions or taking autonomous and even moral choices.[19] It is a questionable habit that survives and conveys false hopes about the current status of AI research. Recently, in a discussion about machine consciousness,[1] it has been claimed that not only research-grade robots, but even Legobots used in first-year undergraduate robot courses are able to develop new motivations. If this were true, why aren't we surrounded by autonomous machines developing their own agenda in order to deal with the environment?

Such approximate misevaluations of the real status of AI hinder researchers from addressing the allegedly, but mistakenly, achieved aforementioned objectives. In the past, many AI researchers made bold claims about their achievements so to endorse a false feeling about the effective level of AI research, which is actually far from reaching the goal of a robot having the slightest form of consciousness.

We offered a comprehensive yet sketchy view of the theoretical landscapes of machine consciousness. As it should be clear, it is a very broad field that stretches significantly the traditional ground for the mind-body problem. It is at the same time a technological and a theoretical field. It compels to address old and new problems using novel and original approaches. We believe machine consciousness will push many researchers to reconsider threads left loose by AI and cognitive science.

Acknowledgments

The authors would like to thank Salvatore Gaglio and Irene Macaluso for useful comments on previous versions of the paper.

References

1. I. Aleksander, U. Awret, S. Bringsjord, R. Chrisley, R. Clowes, J. Parthemore, S. Stuart, S. Torrance and T. Ziemke, Assessing artificial consciousness, *J. Consci. Stud.* **15** (2008) 95–110.
2. I. Aleksander, Machine consciousness, *Scholarpedia* **3** (2008) 4162.
3. C. Adami, What do robots dreams of? *Science* **314** (2008) 1093–1094.
4. G. Buttazzo, Artificial consciousness: Utopia or real possibility, *Spectrum IEEE Comput.* **34** (2001) 24–30.
5. G. Buttazzo, Artificial consciousness: Hazardous questions (and answers), *Artificial Intelligence in Medicine* **44** (2008) 139–146.
6. R. Chrisley, Philosophical foundations of artificial consciousness, *Artificial Intelligence in Medicine* **44** (2008) 119–137.
7. R. Manzotti and V. Tagliasco, Artificial consciousness: A discipline between technological and theoretical obstacles, *Artificial Intelligence in Medicine* **44** (2008) 105–118.
8. A. Chella and R. Manzotti (eds.), *Artificial Consciousness* (Imprint Academic, Exeter (UK), 2007).
9. O. Holland, *Machine Consciousness* (Imprint Academic, New York, 2003).
10. O. Holland, The future of embodied artificial intelligence: Machine consciousness? in *Embodied Artificial Intelligence*, ed. F. Iida (Springer, Berlin, 2004), pp. 37–53.
11. H. Ebbinghaus, *Abriss der Psychologie* (Von Veit, Leipzig, 1908).
12. T. Nemes, *Kibernetic Gépek* (Akadémiai Kiadò, Budapest, 1962).
13. H. Putnam, Robots: Machines or artificially created life? *J. Philosophy* **61** (1964) 668–691.
14. C. Koch, *The Quest for Consciousness: A Neurobiological Approach* (Roberts & Company Publishers, Englewood (Col), 2004).

15. C. Jennings, In search of consciousness, *Nature Neuroscience* **3** (2000) 1.
16. A. Seth, Z. Dienes, A. Cleeremans, M. Overgaard and L. Pessoa, Measuring consciousness: Relating behavioral out neurophysiological approaches, *Trends in Cognitive Sciences* **12** (2008) 314–321.
17. G. Miller, What is the biological basis of consciousness? *Science* **309** (2005) 79.
18. W. G. Lycan, Form, function, and feel, *J. Philosophy* **78** (1981) 24–50.
19. W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, New York, 2008).
20. M. P. Shanahan, Global access, embodiment, and the conscious subject, *J. Consci. Stud.* **12** (2005) 46–66.
21. R. Pfeifer, M. Lungarella and I. Fumiya, Self-organization, embodiment, and biologically inspired robotics, *Science* **5853** (2007) 1088–1093.
22. N. Hirose, An ecological approach to embodiment and cognition, *Cognitive Systems Research* **3** (2002) 289–299.
23. R. Chrisley, Non-conceptual content and robotics: Taking embodiment seriously, in *Android Epistemology*, eds. K. M. Ford, C. Glymour and P. Hayes (AAAI/MIT Press, Cambridge, 1995), pp. 141–166.
24. F. Dretske, Machines and the mental, *Proceedings and Addresses of the American Philosophical Association* **59** (1985) 23–33.
25. B. J. Baars, *A Cognitive Theory of Consciousness* (Cambridge University Press, Cambridge, 1988).
26. R. Sanz, I. Lopez and J. Bermejo-Alonso, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, Exeter (UK), 2007), pp. 141–155.
27. I. Aleksander and B. Dunmall, Axioms and tests for the presence of minimal consciousness in agents, *J. Consci. Stud.* **10** (2003) 7–18.
28. D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, New York, 1996), p. xvii, 414.
29. G. Strawson, *Mental Reality* (MIT Press, Cambridge, 1994), p. xiv, 337.
30. R. Manzotti, Towards artificial consciousness, *APA Newsletter on Philosophy and Computers* **07** (2007) 12–15.
31. C. Koch and G. Tononi, Can machines be conscious? *IEEE Spectrum*, 2008, pp. 47–51.
32. N. Block, On a confusion about a function of consciousness, *Behavioral and Brain Sciences* **18** (1995) 227–287.
33. G. Tononi, An information integration theory of consciousness, *BMC Neuroscience* **5** (2004) 1–22.
34. B. Webb, Can robots make good model of biological behavior? *Behavioral and Brain Sciences* **24** (2001) 1081–1087.
35. F. J. Varela, E. Thompson and E. Rosh, *The Embodied Mind: Cognitive Science and Human Experience* (MIT Press, Cambridge, 1991/1993).
36. A. Clark, *Being There: Putting Brain, Body and World Together Again* (MIT Press, Cambridge, 1997).
37. R. Pfeifer and J. Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence* (Bradford Books, New York, 2006).
38. T. Ziemke and N. Sharkey, A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life, *Semiotica* **134** (2001) 701–746.
39. R. Pfeifer, *Understanding Intelligence* (MIT Press, Cambridge, 1999/2001).
40. R. Manzotti and V. Tagliasco, From "behavior-based" robots to "motivations-based" robots, *Robotics and Autonomous Systems* **51** (2005) 175–190.

41. T. Metzinger, *Being No One: The Self-Model Theory of Subjectivity* (MIT Press, Cambridge, 2003), p. xii, 699.
42. R. Grush, The emulation theory of representation: Motor control, imagery, and perception, *Behavioral and Brain Sciences* **27** (2004) 377–442.
43. S. Lehar, Gestalt isomorphism and the primacy of subjective conscious experience: A Gestalt bubble model, *Behavioral and Brain Sciences* **26** (2003) 375–444.
44. C. Paul, F. J. Valero-Cuevas and H. Lipson, Design and control of tensegrity robots, *IEEE Trans. Robotics* **22** (2006) 944–957.
45. R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati and M. Williamson, The Cog project: Building a humanoid robot, in *Computation for Metaphors, Analogy, and Agents*, ed. C. Nehaniv, Vol. 1562 (Springer-Verlag, Berlin, 1999), pp. 52–87.
46. G. Metta and P. Fitzpatrick, Early integration of vision and manipulation, *Adaptive Behavior* **11** (2003) 109–128.
47. S. Collins, M. Wisse and A. Ruina, A three-dimensional passive-dynamic walking robot with two legs and knees, *Int. J. Robotics Research* **20** (2001) 607–615.
48. R. A. Brooks, New approaches to robotics, *Science* **253** (1991) 1227–1232.
49. J. Bongard, V. Zykov and H. Lipson, Resilient machines through continuous self-modeling, *Science* **314** (2006) 1118–1121.
50. G. Metta, G. Sandini and J. Konczak, A developmental approach to visually guided reaching in artificial systems, *Neural Networks* **12** (1999) 1413–1427.
51. J. Zlatev, The epigenesis of meaning in human beings, and possibly in robots, *Minds and Machines* **11** (2001) 155–195.
52. A. R. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt Brace, New York, 1999).
53. T. Ziemke, On the role of emotion in biological and robotic autonomy, *BiosSystems* **91** (2008) 401–408.
54. M. Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind* (Simon & Schuster, New York, 2006).
55. M. A. Arbib and J. M. Fellous, Emotions: From brain to robot, *Trends in Cognitive Sciences* **8** (2004) 554–561.
56. R. Trappl, P. Petta and S. Payr, *Emotions in Humans and Artifacts* (MIT Press, Cambridge, 2003).
57. J. M. Fellous and M. A. Arbib, *Who Needs Emotions? The Brain Meets the Robots* (Oxford University Press, Oxford, 2003).
58. C. Breazeal, Emotion and sociable humanoid robots, *Int. J. Human Computer Studies* **59** (2003).
59. R. C. Arkin, Moving up the food chain: Motivation and emotion in behavior-based robots, in *Who Needs Emotions? The Brain Meets the Robots*, eds. J. M. Fellous and M. A. Arbib (Oxford University Press, Oxford, 2003), pp. 35–84.
60. R. Manzotti, in *Emotions and Learning in a Developing Robot*, Emotions, Qualia and Consciousness, Casamicciola, Napoli (Italy), Filosofici, I. I. p. g. S., ed. (World Scientific, Casamicciola, Napoli (Italy), 1998), pp. 483–488.
61. P. M. Simons, *Parts: A Study in Ontology* (Clarendon Press, Oxford, 1987).
62. T. Merrick, *Objects and Persons* (Clarendon Press, Oxford, 2001).
63. A. Revonsuo, Binding and the phenomenal unity of consciousness, *Consciousness and Cognition* **8** (1999) 173–185.
64. S. L. Hurley, Action, the unity of consciousness, and vehicle externalism, in *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans (Oxford University Press, Oxford, 2003).

65. D. C. Dennett, *Consciousness Explained*, 1st edn. (Little Brown and Co., Boston, 1991), p. xiii, 511.
66. B. Libet, *Mind Time: The Temporal Factor in Consciousness* (Harvard University Press, Cambridge, 2004).
67. H. Minkowsky, in *Raum und Zeit* (Versammlung Deutscher Naturforscher, Köln Vortrag, Köln, 1908).
68. R. Kane, *The Oxford Handbook of Free Will* (Oxford University Press, New York, 2001).
69. J. Kim, *Supervenience and Mind* (Cambridge University Press, Cambridge, 1993).
70. J. Kim, *Mind in a Physical World* (MIT Press, Cambridge, 1998).
71. J. R. Searle, *The Rediscovery of the Mind* (MIT Press, Cambridge, 1992), p. xv, 270.
72. B. Libet, Unconscious cerebral initiative and the role of conscious will in voluntary action, *Behavioral and Brain Sciences* **8** (1985) 529−566.
73. F. Dretske, *Naturalizing the Mind* (MIT Press, Cambridge, 1995).
74. R. G. Millikan, *Language, Thought and other Biological Categories: New Foundations for Realism* (MIT Press, Cambridge, 1984).
75. M. Tye, Representationalism and the transparency of experience, *Nous* **36** (2002) 137−151.
76. J. A. Fodor, *Concepts: Where Cognitive Science Went Wrong* (Oxford University Press, Oxford, 1998).
77. S. Harnad, Grounding symbolic capacity in robotic capacity, in "*Artificial Route*" *to* "*Artificial Intelligence*"*: Building Situated Embodied Agents*, eds. L. Steels and R. A. Brooks (Erlbaum, New York, 1995).
78. S. Harnad, The symbol grounding problem, *Physica D* (1990) 335−346.
79. M. R. Bennett and P. M. S. Hacker, *Philosophical Foundations of Neuroscience* (Blackwell, Malden, 2003).
80. S. Harnad and P. Scherzer, First, scale up to the robotic turing test, then worry about feem, *Artificial Intelligence in Medicine* **44** (2008) 83−89.
81. G. Galilei, *The Assayer*, 1623.
82. A. S. Eddington, *The Nature of the Physical World* (MacMillan, New York, 1929/1935) p. 361.
83. D. Bohm, A new theory of the relationship of mind and matter, *Philosophical Psychology* **3** (1990) 271−286.
84. G. Strawson, Does physicalism entail panpsychism? *J. Consci. Stud.* **13** (2006) 3−31.
85. W. James, A world of pure experience, *J. Philosophy* **1** (1905) 533−561.
86. E. Mach, *The Analysis of the Sensations* (Dover Publications, New York, 1886/1959).
87. R. Manzotti, An alternative process view of conscious perception, *J. Consci. Stud.* **13** (2006) 45−79.
88. D. Skrbina, *Mind that Abides* (John Benjamins Pub. Dordrecht, 2009).
89. P. O. Haikonen, *The Cognitive Approach to Conscious Machine* (Imprint Academic, London, 2003).
90. A. Nöe, *Action in Perception* (MIT Press, Cambridge, 2004).
91. D. C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, 1st edn. (Bradford Books, Montgomery, 1978), p. xxii, 353.
92. Z. W. Pylyshyn, *Seeing and Visualizing: It's Not What You Think* (MIT Press, Cambridge, 2003).
93. A. N. Whitehead, *Science and the Modern World* (Free Press, New York, 1925).
94. J. Kim, Dretske's qualia externalism, *Philosophical Issues* **7** (1995) 159−165.
95. T. Honderich, Radical externalism, *J. Consci. Stud.* **13** (2006) 3−13.
96. T. Rockwell, *Neither Ghost nor Brain* (MIT Press, Cambridge, 2005).
97. A. Clark, *Supersizing the Mind* (Oxford University Press, Oxford, 2008).

98. K. O'Regan and A. Nöe, A sensorimotor account of visual perception and consciousness, *Behavioral and Brain Sciences* **24** (2001) 939−1011.
99. J. A. Fodor, *The Modularity of Mind: An Essay on Faculty Psychology* (MIT Press, Cambridge, 1983).
100. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice-Hall, New York, 2003).