Machine Consciousness: A Modern Approach

Riccardo Manzotti
IULM University, Italy
corresponding author: Riccardo.manzotti@iulm.it

Abstract

The purpose of the tutorial is to offer an overview of the theoretical and empirical issues in artificial consciousness. Is it possible to devise, project and build a conscious machine? What are the theoretical challenges? What are the technical difficulties? What is the relation between cognition and consciousness? The most promising models will be sketched and contrasted: Global Work Space, Tononi's information integration, Embodied Cognition, Externalist approaches, bio-inspired cognitive architectures. Questions to be discussed include: What advantage does consciousness provide? Is conscious experience the hallmark of a special style of information processing? Would conscious machines be able to outperform intelligent machines?

1. Is consciousness Relevant for AI?

Since 1949 – when Shannon and Weaver cast the foundation for the forthcoming information age (Shannon and Weaver 1949) – computer science, cognitive science, AI and engineering have aimed to replicate the cognitive and mental capabilities of biological beings. To this purpose, various strategies have been envisaged. By and large, we may distinguish various approaches: the symbolic and logical approach of classic AI (Haugeland 1985a; Russell and Norvig 2003), the sensori-motor approach (Pfeifer 1999), neural-network oriented design (Sporns 2011), the bioinspired strategy (Pfeifer, Lungarella et al. 2007b), and the classic AI approach (Russell and Norvig 2003). All these approaches share something – they focus mostly on the intelligent behavior showed by agents. They try to replicate the capability to react to the environment stimuli and to choose the appropriate course of actions. However, something may be missing. According to Russel & Norvig (2003) one of the main goal of AI has been that of designing system that think … "*machine with minds* in the full and literal sense" (Haugeland 1985b). A full-fledged mind inevitably raises the issue of consciousness.

If we take the human being as the target of our efforts, we are immediately struck by something that AI so far has not addressed properly, namely consciousness.

Human beings not only act and behave. They are conscious of what they do and perceive. Somehow, human beings *feel* what happens to them, a condition usually defined as *being conscious* or as having *consciousness*. There is something that it like to be a certain human being (Nagel 1974). Furthermore, there seems to be some strong dependence between autonomy and consciousness.

The problem of consciousness appears so difficult that it has been dubbed the *hard* problem (Chalmers 1996), to the extent that some scientists and philosophers have even argued that it may lie beyond our cognitive grasp (McGinn 1989; Harnad 2003).

For one, there is a crucial question of paramount importance in neuroscience and AI: does consciousness provide a better way to cope with the environment? Or, to put it differently, has consciousness any selective advantage?

At this point and very broadly, there are two conflicting positions. On the one hand, there are authors that set aside consciousness as a philosophical issue of no concern for AI, Cognitive Science and Neuroscience. As Ronald Arkin put it, "Most roboticists are more than happy to leave these debates on consciousness to those with more philosophical leanings" (Arkin 1998). Either because consciousness has no practical consequences or because it is a false problem, these group of authors prefer to focus on more defined issues (vision, problem solving, knowledge representation, planning, learning, language processing). For them, either consciousness is a free bonus at the end of the AI lunch, or is nothing but a by-product of biological/computational processes.

On the other hand, an increasing number of scientists are taking seriously into consideration the possibility that human beings' consciousness is more than an epiphenomenal by-product. Consciousness may be the expression of some fundamental architectural principle exploited by our brain. If this insight were true, it would mean that, in order to replicate human level of intelligence, we ought to tackle with consciousness too.

In support of pro-consciousness group, there is the fact that we have a first-person experience of being conscious, which is not deniable by any amount of theoretical reasoning. In other words, when I feel a pain in my arm, there is something more than the triggering of some appropriate behavioral response. If this feeling had no practical consequences, it would follow that consciousness is epiphenomenal – namely that it has no practical consequences whatsoever. More bluntly, it would follow that consciousness is a useless phenomenon. Such a conclusion would contradict the principle of natural selection – it does not seem likely. Furthermore, in the animal kingdom, there seems to be a correlation between highly adaptable cognitive systems (such as human beings, primates, and mammals) and consciousness. Insects, worms, arthropods, and the like that are usually considered devoid of consciousness are much less adaptable (they are adaptable as a species but not very much as individuals).

As a result, many scientists are now looking for something that explicitly addresses the issue of machine consciousness (Buttazzo 2001; Holland 2003; Holland 2004; Adami 2006; Chella and Manzotti 2007; Aleksander 2008; Aleksander, Awret et al. 2008a; Buttazzo 2008; Chrisley 2008; Manzotti and Tagliasco 2008). So far, there is no accepted consensus as to what consciousness may be. There are several and often conflicting hypotheses. According to some authors, consciousness is the result of a special kind of information process related with information integration (Tononi 2004b; Tononi 2008). According to another group depend on goal generation and development (Manzotti and Tagliasco 2005b), or embodiment (Holland 2004), or a certain kind of information processing akin to the global workspace (Shanahan 2005a; Shanahan 2010), or the replication of imagination and synthetic phenomenology (Aleksander, Awret et al. 2008b; Chrisley 2009b), or emotions (Ziemke 2008a), and so forth.

Furthermore, consciousness is a not only a technical challenge but also a theoretical feat. In this paper, I would like to address two lines of enquiry. On the one hand, I would like to consider and to list a series of fundamental scientific problems that consciousness research cannot set aside. On the other hand, I would like to consider a series of approaches and I will briefly evaluate their pros and cons.

Among the main scientific issues, I would list:

- Cognitive unity
- Intentionality
- Representation
- Freedom
- Temporal integration
- Feeling vs. functing

These issues are paramount both because they are correlated with conscious experience and because they poses a formidable obstacle to our scientific understanding of the nature of consciousness.

In short, the issue of consciousness is still controversial and full of obstacles. We do not yet know how to tackle with it nor how to measure our success. On this regard, Jaegwon Kim stated that (Kim 1998)

> we are not capable of designing, through theoretical reasoning, a wholly new kind of structure that we can predict will be conscious; I don't think we even know how to begin; or indeed how to measure our success.

Yet it may be a necessary step in devising and building a true autonomous and efficient intelligence machine – a machine with a mind. After all, the lack of a formal definition is not necessarily an obstacle that prevents any progress (Koch 2004):

> Historically, significant scientific progress has commonly been achieved in the absence of formal definitions. For instance, the phenomenological laws of electrical current flow were formulated by Ohm, Ampère, and Volta well before the discovery of the electron in 1892 by Thompson. For the time being, therefore, I adopt [a] working definition of consciousness and will see how far I can get with it.

2. Strong and Weak Machine Consciousness

The recent upsurge of interest and optimism as to the possibility of modeling and implementing conscious machines or conscious agents (Buttazzo 2001; Holland 2003; Holland 2004; Adami 2006; Chella and Manzotti 2007; Aleksander 2008; Aleksander, Awret et al. 2008a; Buttazzo 2008; Chrisley 2008; Manzotti and Tagliasco 2008) should not lead anyone to underestimate the many critical issues lurking in the background.

Machine consciousness is not simply a technological issue, but rather a field that poses old unanswered questions such as the relation between information and meaning, the nature of teleology, the unity of the self, the nature of phenomenal experience, and many others. Like psychology, it can be observed that machine consciousness has a long past and a very brief history (Ebbinghaus 1908). Although the term is fairly recent (first time Nemes 1962), the problem has been addressed since Leibniz's mill. 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 (Putnam 1964, p. 669)

> 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.

Machine consciousness is a promising field of enquiry for at least two reasons. First, it assumes that consciousness is a real phenomenon affecting behavior (Jennings 2000; Koch 2004; Miller 2005; Seth, Dienes et al. 2008). Secondly, it suggests the possibility to reproduce, by means of machines, the most intimate aspect of our mind – namely conscious experience. Although many argued against the possibility of machine consciousness mostly because of a priori assumptions ("no machine will ever be like a man"), no one has conclusively argued against such a possibility so far. Biological chauvinism does not seem move from convincing arguments.

Besides, any argument that seems to deny the possibility of machine consciousness is faulty insofar as the same argument would deny the very possibility of human consciousness. For instance, a naïve adversary of machine consciousness may argue that since CPUs and computer memory do not seem to be the right kind of stuff to harbor phenomenal experience, a computer will never be conscious. And yet, borrowing Lycan's words if such (Lycan 1981, p. 37-38)

> 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.

Contrary to classic AI and functionalism, machine consciousness enthusiasts seem to consider that the classic functional view of the mind in terms of either functions or modules (a la Dennett, so to speak) is insufficient to grasp the full scope and capacity of a conscious agent. Therefore, the traditional arguments against strong AI – for instance, Searle's Chinese Room or Block's Chinese nation argument – loose some of their strength. A machine is not necessarily a Turing machine. In fact, although most available machines are instantiations of von Neumann's blue print, other architectures are becoming available. There is no a priori reason why a machine has to be an instantiation of a Turing machine. Views – such as embodiment, situatedness, and externalism – challenge the classic AI disembodied view of a syntactical symbol-crunching machine (Chrisley 1995; Hirose 2002; Shanahan 2005b; Pfeifer, Lungarella et al. 2007a).

Roughly speaking, machines consciousness lies in the middle between biological chauvinism (only brains are conscious) and liberal functionalism (any functional systems behaviorally equivalent is conscious). Its proponents maintain that biological chauvinism could be too narrow and yet they concede that some kind of physical constraints is unavoidable (no multiple realizability).

Recently, many authors emphasized the alleged behavioural role of consciousness (Baars 1988; Aleksander and Dunmall 2003; Sanz 2005; Shanahan 2005b) in an attempt to avoid the problem of phenomenal experience.

Owen Holland suggested that it is possible to distinguish Weak Artificial Consciousness from Strong Artificial Consciousness (Holland 2003). The former approach deals with agents that behave as if they were conscious, at least in some respects. Such view does not need any commitment to the hard problem of consciousness. On the contrary, the latter approach deals with the possibility of designing and implementing agents capable of real conscious feelings.

Although the distinction between weak and strong artificial consciousness sets a useful temporary working ground, it may suggest a misleading view. Setting aside the crucial feature of the human mind – namely phenomenal consciousness – may divert from the understanding of the cognitive structure of a conscious machine. Skipping the so-called "hard problem" could not be a viable option in the business of making conscious machines.

The distinction between weak and strong artificial consciousness is questionable since it is not matched by a mirror dichotomy between true conscious agents and "as if" conscious agents. Yet, human beings are conscious and there is evidence that most animals exhibiting behavioural signs of consciousness are phenomenally conscious. It is a fact that human beings have phenomenal consciousness. They have phenomenal experiences of 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. Finally, they experience being a self with a certain degree of unity. Human consciousness entails phenomenal consciousness at all levels.

In sum, as mentioned above, it would be very bizarre whether natural selection had selected consciousness without any selective advantage. Thus, we cannot but wonder whether it could be possible to design a conscious machine without dealing squarely with the hard problem of phenomenal consciousness. If natural selection went for it, we strongly doubt that engineers could avoid doing the same. Hence it is possible that the dichotomy between phenomenal and access consciousness – and symmetrically the separation between weak and strong artificial consciousness – is eventually fictitious.

While some authors adopted an open approach that does not rule out the possibility of actual phenomenal states in current or future artificial agents (Chella and Manzotti 2007; Aleksander, Awret et al. 2008a), other authors (Manzotti 2007; Koch and Tononi 2008) maintained that a conscious machine is necessarily a phenomenally conscious machine. For them to be conscious is necessarily having phenomenal experiences or having P-consciousness (Block 1995). For instance, Giulio Tononi suggested that the kind of information integration necessary to exhibit a human level of cognitive autonomy is associated to the emergence of consciousness (Tononi 2004a).

3. Scientific Issues

This paragraph will sketch the scientific, theoretical and philosophical issues at the roots of machine consciousness (indeed often of consciousness itself). Too often researchers accept assumptions that are very far from being justified either empirically or theoretically. As a result, many years have been wasted in pursuing goals on the basis on unwarranted premises.

For one, there is no reason why consciousness should be related to biology. So far, no one has ever been able to suggest any kind of necessary link between the carbon-based molecules featured by living organisms and consciousness. For instance, 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. And yet machine consciousness poses many scientific issues that are worth of attention. I will briefly consider each of them.

A. Embodiment

A much heralded crucial aspect of agency has been embodied cognition (Varela, Thompson et al. 1991/1993; Clark 1997; Ziemke and Sharkey 2001; Pfeifer and Bongard 2006). It cannot be in any way underestimated the importance of the interface between an agent 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.

Although we believe that a body is indeed a necessary condition, we wonder whether there had been any clear understanding of embodiment.

Apart from intuitive cases, when is an agent truly embedded? In some sense, there is no such a thing as a not embodied agent, since even the classic AI algorithm has to be implemented as a physical set of instructions running inside a physical device. In some other sense, even a complex robot such as ASIMO is not embodied. It has a very centralized inner controller computing everything. There are many examples of biological agents that would apparently score very well as to embodiment and that do not seem good candidate for consciousness. Take insects, for instance. They show impressive morphological structure that allows them to perform outstandingly well without a very sophisticated cognitive capability.

The notion of embodiment is probably 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 an agent, its body, and its environment.

B. Situatedness

Besides having a body, a conscious agent could need also being part of a real environment. Yet this is controversial. For instance, many authors argued that consciousness could be a purely virtual inner world created inside a system that, to all respects, could avoid any true contact with a real world (Lehar 2003; Metzinger 2003; Grush 2004). They seem to advocate the possibility of a conscious brain in a vat. Yet we have no empirical evidence that an isolated brain would ever be conscious. There are no known real cases. To this extent, the possibility of a pure virtual phenomenal experience is bizarre, and this bizarreness dims its appeal considerably.

If a consciousness requires embodiment and situatedness, a definition of situatedness would be necessary.

Usually, alleged embodied robots such as Brook's agents, Babybot, Passive walkers, and similar (Brooks, Breazeal et al. 1999; Collins, Wisse et al. 2001; Metta and Fitzpatrick 2003; Paul, Valero-Cuevas et al. 2006) 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 could be a unwarranted premise.

True situatedness may involve some kind of developmental integration with the environment such that the behavioral and teleological structure of the agent is the result of 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.

Another fruitful approach is represented by those implementations that outsource part of the cognitive processes to the environment (Brooks 1991). For instance, the field of epigenetic robotics is strongly interested in designing robots capable of developing accordingly with the environment (Metta, Sandini et al. 1999; Zlatev 2001; Bongard, Zykov et al. 2006).

C. Emotions and motivations

It has been maintained that emotions are key to consciousness. For instance, Damasio suggested that there is a core consciousness supporting the higher forms of cognition (Damasio 1999). Although this is a fascinating hypothesis, it remains unclear how emotions should be implemented. Many roboticists draw inspiration from various emotional models (Manzotti 1998; Arkin 2003; Breazeal 2003; Fellous and Arbib 2003; Trappl, Petta et al. 2003; Arbib and Fellous 2004; Minsky 2006; Ziemke 2008b). However, in which case an architecture is really equipped with emotion? When are emotions more than labels on cognitive modules?

Furthermore, it may be the case that emotions depends on consciousness. Another misleading approach has been that offered by the ubiquitous Kismet often described as a robot with emotions (Breazeal 2003). Kismet has nothing to do with emotions apart 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 behaviors for 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.

D. Unity and causal integration

Consciousness seems to depend on the notion of unity. Yet what does it give unity to a collection of parts, being them events, parts, processes, computations, instructions? The ontological analysis has not gone very far (Simons 1987; Merrick 2001) and neuroscience wonders at the mystery of neural integration (Revonsuo 1999; Hurley 2003). Machine consciousness has to face the issue of unity. Would be enough to provide a robot with a series of capabilities for the emergence of a unified agent? Should we consider the necessity of a central locus of processing or the unity would stem out of further unexpected aspects?

Classic theories of consciousness are often vague as to what gives unity to a scattered collection of processes. For instance, would the Pandemonium like community of software demons championed by Dennett (Dennett 1991) become a whole? Has software unity out of its programmer's head? Would embodiment and situatedness be helpful?

A novel approach to the problem of unity is the notion of integrated information introduced by Giulio Tononi (Tononi 2004a). 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 are entangled together. Although still in its initial stage, Tononi's approach may cast a new light on the notion of unity in an agent.

E. Time, duration or present

Conscious experience is located in time. Human beings experience the flow of time in a characteristic way that 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. Surprisingly, there is evidence that half a second of continuous nervous activity is necessary in order to be visually aware of something (Libet 2004).

Furthermore, the classic Newtonian time fits very loosely with our experience of time. According to Newton, only the instantaneous present is real. Everything had to fit in such Euclidean temporal point. Such a present has no duration. For instance, speed is nothing more than the value of a derivative and can be defined at every instant. We assume to occupy only an ever-shifting width-less temporal point. The Einstein-Minkowsky space-time model expresses this view (Minkowsky 1908) – time is a geometrical 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 most neuroscientists, every conscious process is instantiated by patterns of neural activity extended in time. This apparently innocuous hypothesis hides a possible problem. If a neural activity spans in time (as it has to do so since neural activity consists in trains of temporally distributed spikes), something that takes place in different instants of time has to belong to the same cognitive or conscious process. For instance, what glue together the first and the last spike of a neural activity underpinning the perception of 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 machines, 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 would purposefully slow down the speed of such a computation? Certainly, we may envisage an artificial environment where the same computation runs at an altered time (for instance, we may slow down the internal clock of such a machine). Would the alleged conscious machine have a slowed but otherwise identical conscious experience?

A related issue is the problem of the present. As in the case of brains, what does define 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 properties of computations?

Machine consciousness may require a change in our basic notion of time.

F. Will, freedom, and mental causation

Another issue, which does not math with the standard scientific picture of reality, is the fact that a conscious subject seems capable of a unified and free will. The topic is as huge as a topic can be (for a comprehensive review see Kane 2001). The problem connects with the so-called problem of mental causation and top-down causation. If a subject is nothing more than the micro-particles constituting it (and their state also), all causal powers are drained by the smallest constituents. In other words, you and I can't have a will different from what all the particles constituting us do (Kim 1998). If this were true, there will be no space left for any level apart from the lowest one. All reality would be causally reduced to what happens at the micro-particles level. No top-down causation would be possible and no space would remain for the will.

Yet, we have a strong (although possibly wrong) intuition that human beings are capable of influencing their behavior and thus that conscious will makes a difference in the course of events. Many philosophers defended conscious will efficacy (Searle 1992).

Another threat for free will comes from Benjamin Libet's famous studies that showed that awareness of one's own choices follows neural activity by roughly 300 ms (Libet 1985). Although Libet left open the possibility that our consciousness can veto brain deliberations, 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 depend on the system as a whole, a theory of consciousness should be able to address the relation between wholes and parts.

As to machines, the aforementioned issue is quite difficult. The classic mechanistic approach and several respected design strategies (from the traditional *divide et impera* rule of thumb, to sophisticated object-oriented programming languages) suggested to conceive machines as made of separate and autonomous modules. As a result, machines are expected to be cases of physical systems whereas the parts completely drain the causal power of the system as a whole. From this point of view, machines are completely unsuited to endorse a conscious will.

However, two possible approaches can provide a viable escape route out of this blind alley.

The first approach consists in recent connectionist approaches stressing the kind of connectivity between elementary computational units. According to such approaches, it could be possible to implement network whose behavior would stem out of the integrated information of the system as a whole (Tononi 2004a). In short, machines would not have to be mechanistic, after all.

The other 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 the agent would be the result of all its history as a whole. There would not be separate modules dictating what the agent has to do, but rather the past history as a whole would reflect in every choice (Manzotti and Tagliasco 2005a).

G. Representation

One of the most controversial problem in philosophy of mind is that of representation. How is it possible that something represent something else? We face an apparent insurmountable problem. If the physical world were made only of extensional entities that do not refer to anything, the physical world could not possess any semantics. In fact, nobody knows why subject may have semantics in a physical world. The classic Searle's argument suggests that machines could not have intrinsic intentionality and thus are devoid of semantics. If this were true, machines will never be conscious since they will be only syntactic engines. Unfortunately, at the best of our knowledge, the same arguments would rule out brain semantics, too. Why are brains different from machines? Searle's suggestion that brains have special causal powers has never been too persuasive.

Since it is a fact that we have a conscious representation of the world, it conceivable that we need to reframe our view about the physical world in order to accommodate the apparently impossible fact of representation. All attempts to naturalize semantics, intentionality, and representations (with all the known differences among these terms) either failed or did not succeed enough (Millikan 1984; Dretske 1995; Fodor 1998; Tye 2002). How can symbols been grounded with other facts in the world (Harnad 1990; Harnad 1995)?

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 (Bennett and Hacker 2003). 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 brain semantics? In short, I am suspicious of any argument that seems to apply different criteria in biological and in artificial contexts.

In sum, to address the issue of conscious machines, we need to address the issue of representation avoiding any circularity. What does it change a physical process (or state) into a representation of another physical process (or state)?

H. Feeling vs functing, or quantitative vs qualitative

Finally, the allegedly most conspicuous problem – namely how can a physical system produce subjective qualitative phenomenal content? At sunset, we receive boring light rays on our retinas and we experience glorious symphony of colors. We swallow molecules of various kinds and, as a result, we feel the flavour of a delightful Brunello di Montalcino:

> 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." (Harnad and Scherzer 2008)

Famously, Galileo Galilei suggested that smells, tastes, colors, and sounds are nothing without the body of a conscious subject (Galilei 1623). The subject body allegedly creates phenomenal content in some unknown way. A very deep-rooted assumption is the separation between the domain of subjective phenomenal content and the domain of objective physical events. Such assumption deeply intertwines with the deepest epistemological roots of science itself. It is a dogma that a quantitative third-person perspective oblivious of any qualitative aspect can adequately describe physical reality. 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 (Mach 1886; James 1905; Eddington 1929/1935; Bohm 1990; Manzotti 2006; Strawson 2006).

Whether the mental world is a special construct concocted by some irreproducible feature of most mammals is still an open question. There is neither empirical evidence nor theoretical arguments supporting such a view. In the lack of a better theory, many scholars wonder whether would not be wiser to take into consideration the rather surprising idea that the physical world comprehends also those features that we usually attribute to the mental domain (Skrbina 2009). In short, many suspects that some form either of panpsychism or of pan-experientialism ought to be seriously considered.

In the case of machines, how is it possible to take over the so called *functing* vs. *feeling* divide (Lycan 1981; Harnad and Scherzer 2008)? As far as we know, a machine is nothing more than a collection of interconnected modules each functioning in a certain way. Why the functional activity should transfigure in the feeling of a conscious experience? Yet, as it happened for other issues, the same question may 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. So one possibility remains: it is not a problem of the physical world but rather of our picture of the physical world. We may discount a too simplistic view of the physical world. Machines are part of the same physical world that produced conscious human subjects, thus they could take advantage of the same relevant properties and features.

I. Other issues

It is clear that there is a very long list of correlated issued, which I couldn't adequately address here – 1st person vs 3rd person perspectives, intentionality, qualia, relation between phenomenal content and knowledge, special physical phenomena (usually described by quantum laws), mental imagery, meaning, symbol grounding, and so on. It is also true that, while some of these issues partially overlap with the above mentioned topics, some have their own specificity. In general, all problems share a similar structure with respect to machine consciousness: as long as something seems preventing a machine from being conscious, the same condition would deny a brain to be so. Yet, human beings are conscious and thus we should conclude that there must be some mistake in our assumptions about that conditions that apparently deny the very possibility of a conscious physical system.

4. Current Approaches to Machine Consciousness

Although the field is still in its infancy, a few attempts are worth of some consideration. This chapter does not pretend to provide an exhaustive description of these efforts. However, it will be sufficient to overview the ongoing projects.

A. Autonomy and resilience

A conscious agent is a highly autonomous agent. It is capable of self development, learning, self-observation. Is the opposite true?

According to Sanz, there are three motivations to pursue artificial consciousness (Sanz 2005): 1) implementing and designing machines resembling human beings (cognitive robotics); 2) understanding the nature of consciousness (cognitive science); 3) implementing and designing more efficient control systems. The third goal overlaps with the issue of autonomy. A conscious system has to be able to take choices in total autonomy as to its survival and the achievements of its goals. Many authors believe that consciousness endorses a more robust autonomy, a higher resilience, a more general problem solving capability, reflexivity, and self-awareness.

A conscious agent is thus characterized by a strong autonomy that often leads also to resilience to an often huge range of disturbances and unexpected stimuli. Many authors addressed these aspects trying to focus on the importance of consciousness as a control system. Taylor stressed the relation between attention and consciousness (Taylor 2002; Taylor 2007; Taylor 2009) that will be sketched at greater length below. Sanz *et al.* aims to develop a full-fledged functional account of consciousness (Sanz 2005; Sanz, Lopez et al. 2007; Hernandez, Lopez et al. 2009). According to their view, consciousness necessarily emerges from certain, not excessively complex, circumstances in the dwelling of cognitive agents. Finally, it must be quoted Bongard who is trying to implement resilient machines able to recreate their internal model of themselves (Bongard, Zykov et al. 2006). Though he does not stress the link with consciousness, it has been observer that a self-modeling artificial agents has many common traits with a self-conscious mind (Adami 2006).

B. Phenomenal experience in machines

What about explicitly addressing phenomenal experience in machines? There are two approaches, apparently very different: the first approach tries to mimic the functional structure of a phenomenal space (usually vision). The advantage is that it is possible to build robots that exploit the phenomenal space of human beings. For instance, Chrisley is heralding the notion of synthetic phenomenology as an attempt "either to characterize the phenomenal states possessed, or modeled by, an artifact (such as a robot); or 2) any attempt to use an artifact to help specify phenomenal states". (Chrisley 2009a, p.53) Admittedly, Chrisley does not challenge the hard problem. Rather his theory focuses on the sensori-motor structure of phenomenology. Not so differently, Igor Alexander defended various versions of depictive phenomenology (Aleksander and Dunmall 2003; Aleksander and Morton 2007) that suggest the possibility to tackle from a functional point of view the space of qualia.

Another interesting and related approach is that pursued by Antonio Chella who developed a series of robots aiming to exploit sensorimotor contingencies and externalist inspired frameworks (Chella, Gaglio et al. 2001; Chella, Frixione et al. 2008). An interesting architectural feature is the implementation of a generalized closed loop based on the perceptual space as a whole. In other words, in classic feedback only a few parameters are used to control robot behavior (position, speed, etc.). The idea behind the robot is to match a global prediction of the future perceptual state (for instance by a rendering of the visual image) with the incoming data. The goal is to achieve a tight coupling between robot and environment. According to these models and implementations, the physical correlate of robot phenomenology would not lie in the images internally generated but rather in the causal processes engaged between the robot and the environment (Chella and Manzotti 2009).

C. Self motivations

It is a fact that artificial agents do not develop their own goals and thus it is fair to suspect that there is a strong link between being conscious and developing new goals. Up to now there was a lot of interest as to how to learn achieving a goal in the best possible way, but not too much interest as to how develop a new goal. For instance, in their seminal book on neural network learning processes Richard S. Sutton and Andrew G. Barto stresses that they design agent in order to "learn what to do – how to map situations to actions – so as to maximize a numerical reward signal [the goal] […] All learning agents have explicit goals" (Sutton and Barto 1998, p.3-5). In other words, learning deals with situations in which the agent seeks "how" to achieve a goal despite uncertainty about its environment. Yet the goal is fixed at design time. Nevertheless, there are many situations in which it could be extremely useful to allow the agent to look for "what" has to be achieved – namely, choosing new goals and developing corresponding new motivations. In most robots, goals are defined elsewhere at design time (McFarland and Bosser 1993; Arkin 1998) but, at least, behavior changes according to the interaction with the environment.

Interestingly enough, in recent years various researchers tried to design agents capable of developing new motivations and new goals (Manzotti and Tagliasco 2005; Bongard, Zykov et al. 2006; Pfeifer, Lungarella et al. 2007) and their efforts were often related with machine consciousness.

D. Information integration

A possible and novel approach to this problem is the notion of integrated information introduced by Tononi (Tononi 2004). 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 final stage, Tononi's approach could cast a new light on the notion of unity in an agent. Tononi suggested that the kind of information integration necessary to exhibit the kind of behavioural unity and autonomy of a conscious being is also associated to certain intrinsic causal and computational properties which could be responsible for having phenomenal experience (Tononi 2004).

E. Attention

If consciousness has to play a role in controlling the behaviour of an agent, a mechanism that cannot be overlooked is attention control. Attention seems to play a crucial role in singling out to which part of the world to attend. However, it is yet unclear what is the exact relation between attention and consciousness. Though it seems that there cannot be consciousness without attention (Mack and Rock 1998; Simons 2000), there is not sufficient evidence to support the thesis of the sufficiency of attention to bestow consciousness. However, implementing a model of attention is fruitful since introduces many aspects from control theory that could help in figuring out what are the functional advantages of consciousness. This is of the utmost importance since any explanation of consciousness should be tied down to suitable functional ground truth. A satisfying attention control mechanism could satisfy many of the abovementioned goals of consciousness such as autonomy, information integration, perhaps intentionality.

A promising available model of attention is the CODAM neural network control model of consciousness whose main is to provide a functional account (Taylor and Rogers 2002; Taylor 2003; Taylor 2007). Such model has several advantages since it suggests various ways to speed up the response and the accuracy of the agent.

A main advantage of the CODAM neural network control model is that it provides suggestions as to how the brain could implement it. The central idea is that the functional role of the attention copy signal is endorsed by the corollary discharge of attention movement (which is the reason of the name of the model). The possible neural basis of the CODAM has been addressed at length by Taylor (Taylor 2000; Taylor and Rogers 2002; Taylor 2003; Taylor 2007).

F. Global workspace and other cognitive models

A huge area is represented by cognitive models based on some kind of central control structure – often based on the Global Workspace model (Baars 1984) and other likewise cognitive structure. A well-known examples is is Stan Franklin's IDA whose goal is to mimic many high-level behaviors (mostly cognitive in the symbolic sense) gathering together several functional modules. In IDA's top-down architecture, high-level cognitive functions are explicitly modeled (Franklin 1995; Franklin 2003). They aim at a full functional integration between competing software agencies. However, IDA is essentially a functionalist effort. We maintain that consciousness is something more than information processing – it involves embodiment, situatedness and physical continuity with the environment in a proper causal entanglement.

Consider now Gerard Baars' Global Workspace as it has been implemented by Murray Shanahan (Shanahan and Baars 2005; Shanahan 2006). Shanahan's model addresses explicitly several aspects of conscious experience such as imagination and emotion. Moreover, it addresses the issue of sensory integration and the problem of how information is processed in a centralized workspace. It is an approach that, on the one hand, suggests a specific way to deal with information, on the other hand, endorses internalism to the extent that consciousness is seen as the result of internal organization. Consciousness, in short, is a style of information processing (the bidirectional transfer of information from/to the global workspace) achieved through different means – "conscious information processing is cognitively efficacious because it integrates the results of the brain's massively parallel computational resources" (Shanahan 2006, p. 434). He focuses on implementing a hybrid architecture mixing together the more classic cognitive structure of global workspace with largely not symbolic neural networks.

5. What is AI still missing?

Although AI achieved impressive results (Russell and Norvig 2003), it is always astonishing the degree of overvaluation that many non-experts seem to stick to. In 1985 (!), addressing the Americal Philosophical Association, Fred Drestke was sure that "even the simple robots designed for home amusement talk, see, remember and learn" (Dretske 1985, p. 23). It is not unusual to hear that robots are capable of feeling emotions or taking autonomous and even moral choices (Wallach and Allen 2009). It is a questionable habit that survives and that conveys false hopes about the status of AI research. For instance, in a discussion about machine consciousness, it has been claimed that not even research-grade robots, but rather Legobots used in first-year undergraduate robot instruction should be able to develop new motivations (Aleksander, Awret et al. 2008a, p. 102-103). If this were true, why do not autonomous machines developing their own agenda in order to deal with their environment surround us?

Such approximate misevaluation of the real status of AI hinders new researchers from addressing objectives allegedly but mistakenly assumed as already achieved. Due to various motivations not all of strict scientific nature, 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.

In AI, various misunderstandings hamper most approaches to machine consciousness. I list here the possible methodological mistakes that are specific to the field of machine consciousness.

A. "False goals"

Due to its vagueness and intrinsic difficulty, the issue of consciousness has often downgraded to some more tractable aspect. This is an example of the mereological fallacy that consists in confusing a problem with a part of it. For instance, it is true that often a conscious agent is 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 more or less 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, rather obscure. Along these lines, Giulio Tononi and Cristof Koch argued that consciousness does not require many of the skills that AI researchers strive to emulate in machines (Koch and Tononi 2008, p. 50)

> 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.

The issue is still controversial. Most machine consciousness enthusiasts would probably argue against such view – more prominently those that associate conscious agency with the capacity either to integrate cognitive skills (Baars 1988; Haikonen 2003; Shanahan 2005b) or to be autonomous, resilient, and embodied (Sanz 2005; Bongard, Zykov et al. 2006).

B. Labeling

Very often cognitive scientists, roboticists and AI researchers shows 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 architecture models constitute empirical and theoretical claims that must be justified elsewhere. To use Dennett's terminology they are "explanatory debts that have yet to be discharged" (Dennett 1978).

Even an uncontroversial term such as "neural network" is loaded with vague references to biological assumptions. The very choice of the name endorses a series of expectations. Probably, if neural networks had been introduced under the sober name of "not linear functional approximator", their explanatory power would not have been the target of high expectations.

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

C. Confusion between ontological and explanatory levels

It is easy to accept the existence of multiple levels 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 computer memories? Yet the explanatory power of multiple levels ought not to be confused with their reality. It is well known that such use could be a powerful source of confusion (Bennett and Hacker 2003). The use of language is not innocent.

For instance, are images really *inside* a computer memory? Are values inside computers really symbols or characters or whatever we take them to be? 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 even music and images. We could get even higher and consider the existence of images having a certain content. Yet, are all these levels real or are they just different epistemic perspectives on the same phenomenon?

The trouble is that most of these levels – bits, logical values, numbers, RGB triplets - are properties off a way of thinking about what takes place in our computer; they are not properties of the computer as such. What we think for quite naturally as two *pixels* in an image are nothing but two tensions causally related with what happens on a computer screen. On this Zenon Pylyshyn wrote

> The point here is not that a matrix representation is wrong. It's just that it is neutral with respect to the question of whether it models intrinsic (i.e. architectural) properties of mental images. (Pylyshyn 2003, p. 365)

In short all these levels may be akin to a center of mass, insofar as centers of mass do not exist but are simply epistemic shortcuts to refer to complex mass distributions. 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. Yet, why is it real? It is not easy to resist to the reductionist pull draining out from every level except the physical one.

D. Inside and outside

Finally, where is the mind and its content located? Inside or outside the body of the agent? So far, both options are not entirely satisfactory and thus the debate keeps going on.

On one side, it would be very simple if we could locate consciousness inside the body of the agent and thus inside future conscious machines. However, such view is not convincing since most mental states (very broadly speaking) are about something that appears to be external to the body. Therefore, mental states should somehow address external states of affairs (Putnam 1975; Gertler 2007; Lenay and Steiner 2010) – 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, either the mental world must be inside the agent from the beginning or it must be concocted inside (Fodor 1983; Metzinger 2003). All these positions can broadly be labeled as cases of internalism.

On the other hand, consciousness refer to the external world that could be constitutive either as content or as vehicle. 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 (Honderich 2006; Manzotti 2006; Clark 2008). Not only the content of our experience would lie outside our body, but also the vehicles responsible for consciousness may be totally or partially external to the agent’s body. Such a twist in our perspective about the limit of the agent endorses those views that consider embodiment and situatedness as relevant factors for a conscious machine.

6. Conclusion

I tried to outline the present and foreseeable future state of machine consciousness studies. As it should be clear, machine consciousness is a broad field that stretches and enlarges significantly the traditional ground for mind-body problem discussions. It is both a technological and a theoretical field since it addresses old and new problems using a different approach. Machine consciousness will push many researchers to reconsider some threads left loose by classic AI and cognitive science. It may also be that machine consciousness will succeed in shedding a new light on the thorny issue of consciousness.

References

Adami, C., (2006), "What Do Robots Dreams Of?" in *Science*, 314(58): 1093-1094.

Aleksander, I., (2008), "Machine consciousness" in *Scholarpedia*, 3(2): 4162.

Aleksander, I., U. Awret, et al., (2008a), "Assessing Artificial Consciousness" in *Journal of Consciousness Studies*, 15(7): 95-110.

Aleksander, I., U. Awret, et al., (2008b), "Assessing Artificial Consciousness" in *Journal of Consciousness Studies*, 15: 95-110.

Aleksander, I. and B. Dunmall, (2003), "Axioms and Tests for the Presence of Minimal Consciousness in Agents" in *Journal of Consciousness Studies*, 10: 7-18.

Aleksander, I. and H. Morton, (2007), "Depictive Architectures for Synthetic Phenomenology" in A. Chella and R. Manzotti, Eds, *Artificial Consciousness*, Exeter, Imprint Academic(30-45).

Arbib, M. A. and J. M. Fellous, (2004), "Emotions: from brain to robot" in *Trends in Cognitive Sciences*, 8 (12): 554-561.

Arkin, R. C., (1998), *Behavior-Based Robotics*, Cambridge (Mass), MIT Press.

Arkin, R. C., (2003), "Moving Up the Food Chain: Motivation and Emotion in Behavior-Based Robots" in J. M. Fellous and M. A. Arbib, Eds, *Who needs emotions? The brain meets the robots*, Oxford, Oxford University Press**:** 35-84.

Baars, B. J., (1988), *A Cognitive Theory of Consciousness*, Cambridge, Cambridge University Press.

Bennett, M. R. and P. M. S. Hacker, (2003), *Philosophical Foundations of Neuroscience*, Malden (Mass), Blackwell.

Block, N., (1995), "On a Confusion about a Function of Consciousness" in *Behavioral and Brain Sciences*, 18: 227-287.

Bohm, D., (1990), "A new theory of the relationship of mind and matter" in *Philosophical Psychology*, 3(2): 271-286.

Bongard, J., v. Zykov, et al., (2006), "Resilient Machines Through Continuous Self-Modeling" in *Science*, 314(5802): 1118-1121.

Breazeal, C., (2003), "Emotion and Sociable Humanoid Robots" in *International Journal of Human Computer Studies*, 59.

Brooks, R. A., (1991), "New Approaches to Robotics" in *Science*, 253: 1227-1232.

Brooks, R. A., C. Breazeal, et al., (1999), "The Cog Project: Building a Humanoid Robot" in C. Nehaniv, Ed., *Computation for Metaphors, Analogy, and Agents*, Berlin, Springer-Verlag(1562)**:** 52-87.

Buttazzo, G., (2001), "Artificial Consciousness: Utopia or Real Possibility" in *Spectrum IEEE Computer*, 34(7): 24-30.

Buttazzo, G., (2008), "Artificial Consciousness: Hazardous Questions" in *Journal of Artificial Intelligence and Medicine*(Special Issue on Artificial Consciousness).

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

Chella, A., M. Frixione, et al., (2008), "A Cognitive Architecture for Robot Self-Consciousness" in *Artificial Intelligence in Medicine*(Special Issue of Artificial Consciousness).

Chella, A., S. Gaglio, et al., (2001), "Conceptual representations of actions for autonomous robots" in *Robotics and Autonomous Systems*, 34(4): 251-264.

Chella, A. and R. Manzotti, (2007), *Artificial Consciousness*, Exeter (UK), Imprint Academic.

Chella, A. and R. Manzotti, (2009), "Machine Consciousness: A Manifesto for Robotics" in *International Journal of Machine Consciousness*, 1(1): 33-51.

Chrisley, R., (1995), "Non-conceptual Content and Robotics: Taking Embodiment Seriously" in K. Ford, C. Glymour and P. Hayes, Eds, *Android Epistemology*, Cambridge, AAAI/MIT Press**:** 141-166.

Chrisley, R., (2008), "The philosophical foundations of Artificial Consciousness" in *Journal of Artificial Intelligence and Medicine*(Special Issue on Artificial Consciousness).

Chrisley, R., (2009a), "Synthetic Phenomenology" in *International Journal of Machine Consciousness*, 1(1): 53-70.

Chrisley, R., (2009b), "Synthetic Phenomenology" in *International Journal of Machine Consciousness*, 1: 53-70.

Clark, A., (1997), *Being there: putting brain, body and world together again*, Cambridge (Mass), MIT Press.

Clark, A., (2008), *Supersizing the Mind*, Oxford, Oxford University Press.

Collins, S., M. Wisse, et al., (2001), "A Three-dimensional Passive-dynamic Walking Robot with Two Legs and Knees" in *The International Journal of Robotics Research*, 20(7): 607-615.

Damasio, A. R., (1999), *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, New York, Harcourt Brace.

Dennett, D. C., (1978), *Brainstorms: philosophical essays on mind and psychology*, Montgomery, Bradford Books.

Dennett, D. C., (1991), *Consciousness explained*, Boston, Little Brown and Co.

Dretske, F., (1985), "Machines and the Mental" in *Proceedings and Addresses of the American Philosophical Association* 59(1): 23-33.

Dretske, F., (1995), *Naturalizing the Mind*, Cambridge (Mass), MIT Press.

Ebbinghaus, H., (1908), *Abriss der Psychologie*, Leipzig, Von Veit.

Eddington, A. S., (1929/1935), *The nature of the physical world*, New York, MacMillan.

Fellous, J. M. and M. A. Arbib, (2003), *Who needs emotions? The brain meets the robots*, Oxford, Oxford University Press.

Fodor, J. A., (1983), *The Modularity of Mind. An Essay on Faculty Psychology*, Cambridge (Mass), MIT Press.

Fodor, J. A., (1998), *Concepts. Where Cognitive Science went Wrong*, Oxford, Oxford University Press.

Galilei, G., (1623), *The Assayer*.

Gertler, B., (2007), "Content Externalism and the Epistemic Conception of the Self" in *Philosophical Issues*, 17: 37-56.

Grush, R., (2004), "The emulation theory of representation: Motor control, imagery, and perception" in *Behavioral and Brain Sciences*, 27: 377-442.

Haikonen, P. O., (2003), *The Cognitive Approach to Conscious Machine*, London, Imprint Academic.

Harnad, S., (1990), "The Symbol Grounding Problem" in *Physica*, D(42): 335-346.

Harnad, S., (1995), "Grounding symbolic capacity in robotic capacity" in L. Steels and R. A. Brooks, Eds, *"Artificial Route" to "Artificial Intelligence": Building Situated Embodied Agents*, New York, Erlbaum.

Harnad, S., (2003), "Can a machine be conscious? How?" in *Journal of Consciousness Studies*.

Harnad, S. and P. Scherzer, (2008), "First, Scale Up to the Robotic Turing Test, Then Worry About Feem" in *Journal of Artificial Intelligence and Medicine*(Special Issue on Artificial Consciousness).

Haugeland, J., (1985a), "Artificial Intelligence: The very Idea" in, *Mind Design II*, Cambridge (Mass), MIT Press.

Haugeland, J., (1985b), *Semantic Engines: An introduction to mind design*, Cambridge (Mass), MIT Press.

Hernandez, C., I. Lopez, et al., (2009), "The Operative mind: A Functional, Computational and Modeling Approach to Machine Consciousness" in *International Journal of Machine Consciousness*, 1(1): 83-98.

Hirose, N., (2002), "An ecological approach to embodiment and cognition" in *Cognitive Systems Research*, 3: 289-299.

Holland, O., Ed. (2003), *Machine consciousness*, New York, Imprint Academic.

Holland, O., (2004), "The Future of Embodied Artificial Intelligence: Machine Consciousness?" in F. Iida, Ed., *Embodied Artificial Intelligence*, Berlin, Springer**:** 37-53.

Honderich, T., (2006), "Radical Externalism" in *Journal of Consciousness Studies*, 13(7-8): 3-13.

Hurley, S. L., (2003), "Action, the Unity of Consciousness, and Vehicle Externalism" in A. Cleeremans, Ed., *The Unity of Consciousness: Binding, Integration, and Dissociation*, Oxford, Oxford University Press.

James, W., (1905), "A world of pure experience" in *Journal of Philosophy*, 1: 533-561.

Jennings, C., (2000), "In Search of Consciousness" in *Nature Neuroscience*, 3(8): 1.

Kane, R., Ed. (2001), *The Oxford Handbook of Free Will*, New York, Oxford University Press.

Kim, J., (1998), *Mind in a Physical World*, Cambridge (Mass), MIT Press.

Koch, C., (2004), *The Quest for Consciousness: A Neurobiological Approach*, Englewood (Col), Roberts & Company Publishers.

Koch, C. and G. Tononi, (2008), "Can Machines be Conscious?" in *IEEE Spectrum*: 47-51.

Lehar, S., (2003), "Gestalt Isomorphism and the Primacy of Subjective Conscious Experience: A Gestalt Bubble Model" in *Behavioral and Brain Sciences*, 26(4): 375-444.

Lenay, C. and P. Steiner, (2010), "Beyond the internalism/externalism debate: the constitution of the space of perception" in *Consciousness and Cognition*, 19(4): 938-52.

Libet, B., (1985), "Unconscious Cerebral Initiative and the Role of Conscious Will in Volutary Action" in *Behavioral and Brain Sciences*, VIII: 529-566.

Libet, B., (2004), *Mind Time. The Temporal Factor in Consciousness*, Cambridge (Mass), Harward University Press.

Lycan, W. G., (1981), "Form, Function, and Feel" in *The Journal of Philosophy*, 78(1): 24-50.

Mach, E., (1886), *The Analysis of the Sensations*, New York, Dover Publications.

Manzotti, R., (1998), "Emotions and learning in a developing robot", in *Emotions, Qualia and Consciousness*, Casamicciola, Napoli (Italy), World Scientific.

Manzotti, R., (2006), "An alternative process view of conscious perception" in *Journal of Consciousness Studies*, 13(6): 45-79.

Manzotti, R., (2007), "Towards Artificial Consciousness" in *APA Newsletter on Philosophy and Computers*, 07(1): 12-15.

Manzotti, R. and V. Tagliasco, (2005a), "From "behaviour-based" robots to "motivations-based" robots" in *Robotics and Autonomous Systems*, 51(2-3): 175-190.

Manzotti, R. and V. Tagliasco, (2005b), "From behaviour-based robots to motivation-based robots" in *Robotics and Autonomous Systems*, 51: 175-190.

Manzotti, R. and V. Tagliasco, (2008), "Artificial Consciousness: A Discipline Between Technological and Theoretical Obstacles" in *Journal of Artificial Intelligence and Medicine*(Special Issue on Artificial Consciousness).

McGinn, C., (1989), "Can we Solve the Mind Body Problem?" in *Mind*, 98: 349-366.

Merrick, T., (2001), *Objects and Persons*, Oxford, Oxford Clarendon Press.

Metta, G. and P. Fitzpatrick, (2003), "Early integration of vision and manupulation" in *Adaptive Behavior*, 11(2): 109-128.

Metta, G., G. Sandini, et al., (1999), "A developmental approach to visually guided reaching in artificial systems" in *Neural Networks*, 12: 1413-1427.

Metzinger, T., (2003), *Being no one: the self-model theory of subjectivity*, Cambridge (Mass), MIT Press.

Miller, G., (2005), "What is the Biological Basis of Consciousness?" in *Science*, 309: 79.

Millikan, R. G., (1984), *Language, Thought, and other Biological Categories: New Foundations for Realism*, Cambridge (Mass), MIT Press.

Minkowsky, H., (1908), "Raum und Zeit", in *Versammlung Deutscher Naturforscher*, Köln Vortrag.

Minsky, M., (2006), *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* New York, Simon & Schuster.

Nagel, T., (1974), "What is it like to be a Bat?" in *The Philosophical Review*, 4: 435-450.

Nemes, T., (1962), *Kibernetic Gépek*, Budapest, Akadémiai Kiadò.

Paul, C., F. J. Valero-Cuevas, et al., (2006), "Design and Control of tensegrity Robots" in *IEEE Transactions on Robotics*, 22(5): 944-957.

Pfeifer, R., (1999), *Understanding Intelligence*, Cambridge (Mass), MIT Press.

Pfeifer, R. and J. Bongard, (2006), *How the Body Shapes the Way We Think: A New View of Intelligence (Bradford Books)* New York, Bradford Books.

Pfeifer, R., M. Lungarella, et al., (2007a), "Self-Organization, Embodiment, and Biologically Inspired Robotics" in *Science*, 5853(318): 1088 - 1093.

Pfeifer, R., M. Lungarella, et al., (2007b), "Self-Organization, Embodiment, and Biologically Inspired Robotics" in *Science*, 5853: 1088-1093.

Putnam, H., (1964), "Robots: Machines or Artificially Created Life?" in *The Journal of Philosophy*, 61(21): 668-691.

Putnam, H., (1975), *Mind, language, and reality*, Cambridge, Cambridge University Press.

Pylyshyn, Z. W., (2003), *Seeing and Visualizing. It's Not What You Think*, Cambridge (Mass), MIT Press.

Revonsuo, A., (1999), "Binding and the phenomenal unity of consciousness" in *Consciousness and Cognition*, 8: 173-85.

Russell, S. and P. Norvig, (2003), *Artificial Intelligence. A Modern Approach*, New York, Prentice Hall.

Sanz, R., (2005), "Design and Implementation of an Artificial Conscious Machine", in *IWAC2005*, Agrigento.

Sanz, R., I. Lopez, et al., (2007), "Principles for consciousness in integrated cognitive control" in *Neural Networks*, 20: 938-946.

Searle, J. R., (1992), *The rediscovery of the mind*, Cambridge (Mass), MIT Press.

Seth, A., Z. Dienes, et al., (2008), "Measuring consciousness: relating behavioural out neurophysiological approaches" in *Trends in Cognitive Sciences*, 12(8): 314-321.

Shanahan, M., (2005a), "Global Access, Embodiment, and the Conscious Subject" in *Journal of Consciousness Studies*, 12: 46-66.

Shanahan, M., (2010), *Embodiment and the Inner Life. Cognition and Consciousness in the Space of Possible Minds*, Oxford, Oxford University Press.

Shanahan, M. P., (2005b), "Global Access, Embodiment, and the Conscious Subject" in *Journal of Consciousness Studies*, 12(12): 46-66.

Shannon, C. E. and W. Weaver, (1949), *The Mathematical Theory of Communication*, Urbana, University of Illinois Press.

Simons, P. M., (1987), *Parts. A Study in Ontology*, Oxford, Clarendon Press.

Skrbina, D., Ed. (2009), *Mind that abides*, Dordrecht, John Benjamins Pub.

Sporns, O., (2011), *Networks of the Brain*, Cambridge (Mass), MIT Press.

Strawson, G., (2006), "Does physicalism entail panpsychism?" in *Journal of Consciousness Studies*, 13(10-11): 3-31.

Taylor, J. G., (2002), "Paying attention to consciousness" in *TRENDS in Cognitive Sciences*, 6(5): 206-210.

Taylor, J. G., (2007), "CODAM: A neural network model of consciousness" in *Neural Networks*, 20: 983-992.

Taylor, J. G., (2009), "Beyond Consciousness?" in *International Journal of Machine Consciousness*, 1(1): 11-22.

Tononi, G., (2004a), "An information integration theory of consciousness" in *BMC Neuroscience*, 5(42): 1-22.

Tononi, G., (2004b), "An information integration theory of consciousness" in *BMC Neuroscience*, 5: 1-22.

Tononi, G., (2008), "Consciousness as integrated information: a provisional manifesto." in *Biological Bulletin*, 215: 216-42.

Trappl, R., P. Petta, et al., (2003), *Emotions in Humans and Artifacts*, Cambridge (Mass), MIT Press.

Tye, M., (2002), "Representationalism an the Transparency of Experience" in *Nous*, 36(1): 137-151.

Varela, F. J., E. Thompson, et al., (1991/1993), *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge (Mass), MIT Press.

Wallach, W. and C. Allen, (2009), *Moral Machines. Teaching Robots Right from Wrong*, New York, Oxford University Press.

Ziemke, T., (2008a), "On the role of emotion in biological and robotic autonomy" in *BioSystems*, 91: 401-408.

Ziemke, T., (2008b), "On the role of emotion in biological and robotic autonomy" in *BiosSystems*, 91: 401-408.

Ziemke, T. and N. Sharkey, (2001), "A stroll through the worlds of robots and animals: applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life" in *Semiotica*, 134 (1/4): 701/46.

Zlatev, J., (2001), "The Epigenesis of Meaning in Human Beings, and Possibly in Robots" in *Minds and Machines*, 11: 155-195.