Chapter 20
Artificial Consciousness

Antonio Chella and Riccardo Manzotti

Abstract "Artificial" or "machine" consciousness is the attempt to model and implement aspects of human cognition that are identified with the elusive and controversial phenomenon of consciousness. The chapter reviews the main trends and goals of artificial consciousness research, as environmental coupling, autonomy and resilience, phenomenal experience, semantics or intentionality of the first and second type, information integration, attention. The chapter also proposes a design for a general "consciousness oriented" architecture that addresses many of the discussed research goals. Comparisons with competing approaches are then presented.

20.1 Introduction

Artificial consciousness, sometimes labeled as *machine consciousness*, is the attempt to model and implement those aspects of human cognition which are identified with the often elusive and controversial phenomenon of consciousness (Aleksander 2008; Chella and Manzotti 2009). It does not necessarily try to reproduce human consciousness as such, insofar as human consciousness could be unique due to a complex series of cultural, social, and biological conditions. However, many authors have suggested one or more specific aspects and functions of consciousness that could, at least in principle, be replicated in a machine.

At the beginning of the information era (in the 1950s), there was no clear-cut separation between intelligence and consciousness. Both were considered vaguely overlapping terms referring to what the mind was capable of. For instance, the famous Turing Test was formulated in such a way so as to avoid any commitment about any distinction between a human-like intelligence machine and a human-like conscious machine. As a result, the main counterargument to the Turing Test was raised by a philosopher of the mind (Searle 1980). There was no boundary between

A. Chella (✉)
Department of Computer Engineering, University of Palermo,
Viale delle Scienze – Building 6, 90128 Palermo, Italy
e-mail: chella@unipa.it

V. Cutsuridis et al. (eds.), *Perception-Action Cycle: Models, Architectures, and Hardware*, Springer Series in Cognitive and Neural Systems 1,
DOI 10.1007/978-1-4419-1452-1_20,

intelligence and consciousness. Similarly, most of the cybernetic theory explicitly dealt with the mind as a whole. It is not by chance that the first mention of the term artificial consciousness appears in those years in to a book of cybernetics by Tihamér Nemes, *Kibernetikai gépek* (Nemes 1962), later translated in English (Nemes 1969).

In the following years, in the aftermath of the cybernetic decline, the idea of designing a conscious machine was seldom mentioned because the very notion of consciousness was considered highly suspicious. The reasons for this long lasting scientific banishment are articulated at length in many excellent books (Searle 1992; Chalmers 1996). However, since the beginning of the 1990s, a new scientific interest for consciousness arose (Crick 1994; Hameroff et al. 1996; Miller 2005) leading to current widespread approaches in neuroscience (Jennings 2000; Koch 2004; Adami 2006). Such increased acceptance of the topic allowed many researches in Robotics and AI to reconsider the possibility of modeling and implementing a conscious machine (Buttazzo 2001; Holland 2003, 2004; Adami 2006; Chella and Manzotti 2007; Aleksander 2008; Aleksander et al. 2008; Buttazzo 2008; Chrisley 2008; Manzotti and Tagliasco 2008; Chella and Manzotti 2009).

Before outlining the details of some critical yet promising aspects of artificial consciousness, a preliminary caveat is useful. As we have mentioned, artificial consciousness assumes that there is some aspect of the mind (no ontological commitments here, we could have used the word *cognition* had it not been associated with a conscious-hostile view of the mind) that has not yet been adequately addressed. Therefore, scholars in the field of artificial consciousness suspect that there could be something more going on in the mind than what is currently under scrutiny in field such as artificial intelligence, cognitive science, and computer science. Of course, most AI researchers would agree that there is still a lot of work to do: better algorithms, more data, more complex, and faster learning structures. However, it could be doubted whether this improvement in AI would ever lead to an artificial agent equivalent to a biological mind or it would rather miss some necessary aspect. In the field of artificial consciousness, scholars suspect that AI missed something important.

There are two main reasons that support this suspicion and that encourage an upsurge of interest in artificial consciousness: (1) the gap between artificial and biological agents; (2) the unsatisfactory explanatory power of cognitive science as to certain aspects of the mind such as phenomenal experience and intentionality.

The former problem is how to bridge the still huge chasm dividing biological intelligent and conscious agents from artificial ones – most notably in terms of autonomy, semantic capabilities, intentionality, self-motivations, resilience, and information integration. These are the main problems still waiting to have a solution in engineering terms and it is, at the same time, encouraging and disparaging that conscious agents seem to deal so seamlessly with them.

The latter problem is more theoretical, but endorses many practical issues as well. Not only artificial agents are a poor replica of biological intelligent agents, but – more worryingly – the models derived from artificial intelligence and cognitive science did not succeed in explaining the human and the animal mind. What is the missing ingredient? Could it be either consciousness or something closely linked to it?

For instance, why are not we able to design a working semantic search engine? Yes, we are aware that there is a great deal of interest as to the semantic web and other related project that herald their grasp of semantic aspects of information. Unfortunately, most of the work in the area, apart from its technical brilliance, does not uncover a lot of ground as to what semantics is. For instance, most of semantic engines are simply multiple layered syntactic engines. Instead of storing just the word "Obama," you store also other tags like "president," "USA," and such. While this could provide the impression that the system knows something about who Obama is, there is no semantics inside the system as to what such strings mean.

It is not a secret that the problem of consciousness is hindered by many deep scientific and philosophical conundrums. It is not altogether obvious whether anything like artificial consciousness is possible. After all there could be some physical constraints unbeknownst to us that would prevent a machine without an evolutionary history and a DNA-based constitution to ever express consciousness. Yet, for lack of a strong reason to believe in the a priori impossibility of artificial consciousness, it seems more fruitful to approach it head on.

Before getting into the details of the various aspects of artificial consciousness, it is useful to tackle with a few rather broad question that need to be answered, at least temporarily, in order to be able to flesh out the general structure of the problem.

The first question to consider is whether consciousness is real or not. For many years, the problem has been either simply dismissed or declared to be ill-conceived. Such approach is no longer acceptable mainly because of the wide acceptance that consciousness studies had in the neurosciences (Atkinson et al. 2000; Jennings 2000; Crick and Koch 2003; Miller 2005).

The second question is whether there is any theoretical constraint preventing humans to build a device with a mind comparable to that of a human being. Once more we are confronted more with our prejudices than with either any real empirical evidence or theoretical reason. In 2001, at the Cold Spring Harbour Laboratories (CSHL), a distinguished set of scholars in neuroscience and cognitive science answered negatively to this question by agreeing that "There is no known law of nature that forbids the existence of subjective feelings in artifacts designed or evolved by humans" (quoted in Aleksander 2008) – thereby opening the road to serious attempts at designing and implementing a conscious machine. After all, our brain and body allow the occurrence of consciousness (we have been careful not to say that the brain creates consciousness, we do not want to fall in the mereological fallacy, see Bennett and Hacker 2003).

The third question addresses the issue of the physical or functional nature of consciousness. In other words, is consciousness the result of a particular functional organization or does it emerge out of specific physical phenomena? It could be that the principle of multiple realizability simply does not hold for consciousness: a specific physical process could be required. As simulated water is not wet, a functionally equivalent agent could be consciousness-free whether realized without certain critical components. However, there is no evidence as to what such critical phenomena could be. Up to now all suggested physical phenomena (e.g., the infamous microtubule) did not live up to empirical investigation. On the other hand, if consciousness stems out of functional structures what they could be.

The fourth and final issue is a more specific version of the previous one – if consciousness emerges out of functional states, is it a kind of computation or something else? Is consciousness an algorithmic phenomenon of computational nature? After all, since David Marr, the dominant view is that cognition is computation, if not, what else? It is a view that has been sometimes labeled as symbolic *computationalism*. Famously, Newell stated that "... although a small chance exists that we will see a new paradigmemerge for mind, it seems unlikely to me. Basically, there do not seem to be any viable alternatives [to computationalism]" (Newell 1990, p. 5). Although many argued strongly against this view, there is no consensus as to what the next step could be (Bringsjord 1994; Chrisley 1994; Harnad 1994; Van Gelder 1995): Could artificial consciousness bridge the gap? Time will tell.

Luckily, artificial consciousness does not need to be committed to a particular theoretical view since it can appeal to a real implementation that would hopefully overcome any theoretical limits. This is the invoked strength of an engineering discipline that does not always need to be constrained by our epistemic limits (Tagliasco 2007). So basically, a researcher in the field of artificial consciousness should answer optimistically to all of the previous three questions: (1) consciousness is a real physical phenomenon; (2) consciousness could be replicated by an artificial system designed and implemented by humans; and (3) consciousness is either a computational phenomenon or something more, either way it can be both understood and replicated.

20.2 Goals of Artificial Consciousness

Here we would not outline an implementable model of consciousness since, frankly, such a model is still to be conceived. Besides there are strong doubts as to what be necessary for consciousness. Recently, Giulio Tononi and Cristof Koch argued that consciousness does not require many of the skills that roboticists are trying to implement: "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." (Koch and Tononi 2008, p. 50). Rather we will focus on what an artificial conscious machine should achieve. Too 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 a architecture model constitute empirical and theoretical claims that must be justified elsewhere – at best, they are "explanatory debts that have yet to be discharged" (Dennett 1978).

Roughly speaking, machine consciousness lies in the middle between the two extremes of 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 could be unavoidable (no multiple realizability).

Recently, many authors emphasized the alleged behavioral role of consciousness (Baars 1998; Aleksander and Dunmall 2003; Sanz 2005; Shanahan 2005) in an

attempt to avoid the problem of phenomenal experiences. Owen Holland suggested that it is possible to distinguish Weak Artificial Consciousness from Strong Artificial Consciousness (Holland 2003). The former approach deals with agents which behave as if they were conscious, at least in some respects. Such view does not need any commitment to the hard problem of consciousness thereby suggesting a somehow smoother path to the final target (Seth 2009). On the contrary, strong artificial consciousness deals squarely with the possibility of designing and implementing agents capable of real conscious feelings.

Although the distinction between weak and strong artificial consciousness could set a useful temporary working ground, it could also suggests a misleading view. Setting aside the crucial feature of the human mind – namely phenomenal consciousness – could miss something indispensable for 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 should be 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 behavioral 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, 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 it actually being phenomenally so. Could we really hope to be smarter than natural selection in this respect, sidestepping the issue of phenomenal consciousness? 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 et al. 2008), other authors (Manzotti 2007; Koch and Tononi 2008) maintained that a conscious machine is necessarily a phenomenally conscious machine. Yet, whether having phenomenal consciousness is a requisite or an effect of a unique kind of cognitive architecture is still a shot in the dark.

On the basis of artificial consciousness, a research working program can be outlined. One way to do it is focusing on the main goals that artificial consciousness should achieve: autonomy and resilience, information integration, semantic capabilities, intentionality, and self-motivations. There could be some skepticism as to the criteria to select such goals. There are two main criteria: one is contingent and the

other is more theoretical: (1) the relevance in the artificial consciousness literature; (2) the incapability of traditional cognitive studies to outline a convincing solution.

It is significant that artificial consciousness is not the only game in town challenging the same issues. Other recent approaches were suggested in order to overcome more or less the same problems. An incomplete list includes topics like artificial life, situated cognition, dynamic systems, quantum computation, epigenetic robotics, and many others. It is highly probable that there is a grain of truth in each of these proposals, and it is also meaningful that they share most of their ends if not their means.

Let us consider the main goal of artificial consciousness one by one starting with a rather overarching issue: does consciousness requires a tight coupling with the environment or is it an internal feature of a cognitive system?

20.2.1 Environment Coupling

Under the label of environment coupling we refer to a supposedly common feature of conscious agents: they act in a world they refer to. Their goals are related to facts of the world and their mental states refer to world events. It is perhaps paradoxical that the modern notion of the mind, originated with the metaphysically unextended and immaterial Cartesian soul, is now a fully embodied situated goal-oriented concept.

As we will see, there are various degrees of environment coupling, at least in the theoretical landscape: from embodied cognition to situations, from semantic externalism to radical externalism.

Yet – as much reasonable as the idea that the mind is the result of a causal coupling with the world may seem – it was not the most popular neither for the layman nor for the scientist. The most common doctrine as to the nature of mental phenomena is that consciousness stems out of the brain as an intrinsic and internal property of it. A few years ago, John Searle unabashedly wrote that "Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain" (Searle 1992, p. 1). Some years later, on *Nature*, George Miller stated that "Different aspects of consciousness are probably generated in different brain regions." (Miller 2005, p. 79). Others added that if the problem of consciousness had to shift from a philosophical question to a scientific one resurrecting "a field that is plagued by more philosophical than scientifically sound controversies" (Changeux 2004, p. 603), the only conceivable option is to look inside the nervous system – as if physical reality were restricted to neural activity alone. However, such premises are still empirically undemonstrated and could get overturned by future experimental results. Koch provides a clear statement summing up the core of these widespread beliefs (Koch 2004, p. 16, italics in the original): "The goal is to discover *the minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept*."

This goal is based on the premise that there must be a set of neural events *sufficient* for a specific conscious percept (Crick and Koch 1990). Yet, such a premise

has never been empirically demonstrated so far. Just to be clear, what is at stake is not whether neural activity is necessary but whether neural activity is either *sufficient for or identical with* phenomenal experience. Is consciousness produced *inside* the brain?

By and large, this view does not conflict with the obvious fact that human brains need to develop in a real environment and are the result of their individual history. The historical dependency on development holds for most biological subsystems: for instance, muscles and bones need gravity and exercise in order to develop properly. But, once developed, they are sufficient to deliver their output, so to speak. They need gravity and weight in order to develop, but when ready, muscles are sufficient to produce a variable strength as a result of the contraction of myofibrils. Alternatively, consider the immune system. It needs a contact with the *Varicella Zoster* virus (chicken pox) in order to develop the corresponding antigens. Yet, subsequently, the immune system is sufficient to produce such an output. In short, historical dependence during development is compatible with sufficiency once the system is developed. In the case of consciousness, for most neuroscientists, the environment is necessary for development of neural areas, but it is not constitutive of conscious experience when it occurs in a grown-up normally developed human being. Many believe that consciousness is produced by the nervous system like strength is produced by muscles – the neural activity and consciousness having the same relation as myofibrils and strength.

What has to be stressed is that, although many scientists boldly claim that there is plenty of evidence showing that "the entire brain is clearly sufficient to give rise to consciousness" (Koch 2004, p. 16), actually there is none. The "central plank of modern materialism – the supposition that consciousness supervenes on the brain" (Prinz 2000, p. 425) is surprisingly poorly supported by experimental evidence, up to now.

If the thesis that consciousness is only the result of what is going on inside the brain (or a computer), we must then explain how the external world can contribute to it and thus adopt a more ecological oriented view. The most notable example is offered either by situated or embodied cognition – i.e., the idea that cognition is not only a symbolic crunching performed inside a system, but rather a complex and extended network of causal relation between the agent, its body, and its environment (Varela et al. 1991/1993; Clark 1997; Ziemke and Sharkey 2001; Pfeifer and Bongard 2006; Clark 2008). Although a view very popular, it gained its share of detractors too (for instance, Adams and Aizawa 2008, 2009). It is a view that has been developed in various degrees.

The more conservative option is perhaps simple embodiment – namely, the hypothesis that the body is considered as an integral part of the cognitive activity. Thus, cognitive activity is not constrained to the dedicated information processing hardware, but rather it comprehends the perception–action loops inside the agent body. Such an approach is considered to be advantageous since the body takes into account many physical aspects of cognition that would be difficult to emulate and simulate internally. These approaches were triggered by the seminal work of Rodney Brooks in the 1990s (Brooks 1990, 1991; Brooks et al. 1998, 1999;

Collins et al. 2001; Metta and Fitzpatrick 2003; Paul et al. 2006). However, further efforts have not provided the expected results. In other words, the morphology of the agent seems a successful cognitive structure only when the task is relatively physical and related to bodily action. In short, embodiment is great to take charge of the computational charge of walking, grasping, running, but it is unclear whether it can be useful to develop higher cognitive abilities (Prinz 2009). Sometimes situated cognition is nothing but a more emphasized version of embodiment although it usually stresses the fact that all knowledge is situated in activity bound to social, cultural, and physical contexts (Gallagher 2009).

A more literal version of environmental coupling is developed under the shorthand of externalism in its various versions (Rowlands 2003; Hurley 2006). By and large these positions assume that the perceived environment is not only necessary or useful, but literally constitutive of what the mind is. In other words, these views go beyond both functionalism and embodiment/situatedness. They maintain that the cognitive agent's boundaries have to be extended so to comprehend a relevant part of the environment. This goal can be addressed at various degrees of commitment to the general thesis. To cut a long story short, we can distinguish between cognitive externalism and phenomenal externalism. The former holds that, though cognition is extended in the sense that the agent take advantage of structures which are external to the body of the agent, phenomenal experience as such is still a property of processes occurring inside the agent – the most notable example is perhaps the extended mind model by Clark and Chalmers (1999) though others have defended similar views (Robbins and Aydede 2009). The latter view, labeled as phenomenal externalism, is more daring and takes in consideration the possibility that the vehicles of phenomenal experience are physically larger than the agent body. The conscious mind would thus literally comprehend part of the environment in a strong physical sense (Rockwell 2005; Honderich 2006; Manzotti 2006).

20.2.2 Autonomy and Resilience

A conscious agent is a highly autonomous agent. It is capable of self development, learning, and 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); and (3) implementing and designing more efficient control systems. The third goal is strongly overlapped with the issue of autonomy. A conscious system is expected 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.

It is unquestionable that a conscious agent seems to have a higher adaptability to unpredictable situations. This conflates against the so-called epiphenomenal view

of consciousness according to which consciousness would have no real positive effects on behavior (Libet et al. 1983; Pockett 2004). Yet this view fails when considered the appearance of conscious agents as a result of natural selection – whatever consciousness is, it consumes resources, thus it must have same advantage. Since many evidence does show that most if not all repetitive mental skills can be performed more efficiently in the absence of consciousness (from playing videogames to more complex sensorimotor tasks), it makes sense to suppose that consciousness has a role in keeping together all cognitive processes so as to pursue a common goal.

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, 2007, 2009) that will be sketched at greater length below. Sanz et al. aim to develop a full-fledged functional account of consciousness (Sanz 2005; Sanz et al. 2007; Hernandez 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 et al. 2006). Though he does not stress the link with consciousness, it has been observed that a self-modeling artificial agent has many common traits with a self-conscious mind (Adami 2006).

20.2.3 Phenomenal Experience

This is the most controversial and yet the more specific aspect of consciousness. It is so difficult that in the 1990 had been labeled as the "hard problem" meaning that it entails some very troublesome and deep aspect of reality. Most authors agree that there is no way to sidestep it, and yet there does not seem to be any way to deal with it.

In short, the hard problem is the following. The scientific description of the world seems devoid of any quality of which a conscious agent (a human being) has an experience of. So, in scientific terms, a pain is nothing but a certain configuration of spikes through nerves. However, from the point of view of the agent, the pain is *felt* and not as a configuration of spikes but rather with a very excruciating and unpleasant quality. What is this quality? How it is produced? Where does it take place? Why is it there? "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). Either there is a deep mystery or our epistemic categories are fundamentally flawed. The difficulty we have in tackling with the hard problem of consciousness could indeed be the sign that we need to upgrade our scientific worldview.

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, we wonder whether it would 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 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 but a collection of interconnected modules functioning in a certain way. Why the functional activity of a machine should 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. Nevertheless, we must admit that we could discount a too simplistic view of the physical world.

It is thus possible that a successful approach to the phenomenal aspect of artificial consciousness will not only be a model of a machine, but rather a more general and overarching theory that will deal with the structure of reality.

Given the difficulties of this problem, it is perhaps surprising that many scholars tried to tackle it. 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. Although there in neither explicit nor implicit solution of the hard problem, these attempts highlight that many so-called ineffable features of phenomenal experiences are nothing but functional properties of perceptual space. This is not to say that phenomenal qualities are reducible to functional properties of perceptual spaces. However, these approaches could help to narrow down the difference. 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 2009, p. 53). Admittedly, Chrisley does not challenge the hard problem. Rather his theory focuses on the sensorimotor 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 one of the authors (Chella) who developed a series of robots aiming to exploit sensorimotor contingencies and externalist inspired frameworks (Chella et al. 2001, 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).

Not all authors would consider such approaches satisfying. Most philosophers would argue that unless there is a way to naturalize phenomenal experience it is useless to try to implement it. Although it is a feasible view, it is perhaps worth considering that phenomenal experience in human being is a result of natural selection and that, as such, was not selected as the answer to a theory, but as a solution to one or more practical problems. In this sense, keep trying to model phenomenal experience in robots, albeit with all well-known limitations, may unwittingly help us to understand what phenomenal experience is.

20.2.4 Semantics or Intentionality of the First Type

Semantics is the holy grail of AI symbolic manipulation. It is the missing ingredient of AI. Without semantics it is even doubtful, whether it is meaningful talking of symbols. For, what is a symbol without a meaning? A gear is not a symbol, although is part of a causal network and although it has causal properties that could be expressed in syntactical terms. In fact, all implementations of symbols are causal networks isomorphic to a syntactical space.

Although, since the 1950s, AI and computer science have tried to ground semantics in syntax, it seems conceivable that it ought to be the other way round: syntax could be grounded on semantics.

As we have mentioned, semantics can be seen as a synonym of intentionality of the first type – i.e., intentionality as it was defined by the philosopher Franz Brentano in his seminal work on psychology (Brentano 1874/1973). According to him, intentionality is the arrow of semantics: what links a symbol (or a mental state) to its content whether it be a physical event, a state of affair in the real world, or a mental content. It is not by chance that intentionality of this kind was suggested as the hallmark of the mental.

Brentano's suggestion triggered a long controversy that is still lasting. A famous episode what John Searle's much quoted paper on the Chinese room (Searle 1980). In that paper, he challenged the view that information, syntax, and computation endorse the meaning (the intentionality) of symbols. More recently, the problem of semantics was reframed as the "symbol grounding problem" by Harnad (Harnad 1990).

Not surprisingly, one of the main argument against machine consciousness has been the apparent lack of semantics of artificial systems whose intentionality is allegedly derived from that of their users and designers (Harnad 2003). Human beings (and perhaps animals) have intrinsic intentionality that gives meaning to their brain states. On the contrary, artificial systems like computers or stacks of cards do not have intrinsic intentionality. A variable x stored in my computer is not about my

bank account balance. It is the use that I make of that variable that allows me to attribute it a meaning. However, by itself, that variable has no intrinsic intentionality, no semantics, and no meaning. Or so it goes the classic argument.

Yet, it remains vague both what intrinsic intentionality is and how a human being is able to achieve it. For lack of a better theory, we could envisage to turn the argument upside down. First, we could observe that there are physical systems (human beings) capable of intentionality of the first kind. These systems are conscious agents. Then, we could suspect that, as Brentano suggested, being conscious and having intentionality could be two aspects of the same phenomenon. If this were true, there would be only one way to achieve semantics – namely, designing a conscious agent.

20.2.5 Self-Motivations or Intentionality of the Second Type

Usually artificial agents do not develop their own motivations. Their goals and hierarchy of subgoals are carefully imposed at design time. This is desirable since the goal of artificial agents is to fulfill their designers' goals, not to wander around purchasing unbeknownst goals. Yet, conscious agents like humans seem capable to develop their own goals and, indeed, it seems a mandatory feature of a conscious being. A human being incapable of developing his/her own agenda would indeed seem curiously lacking something of important.

The capability of developing new goals is deeply intertwined with intentionality according to the Dennett's definition of it: having an intention of pursuing a certain state of affairs (Dennett 1987, 1991). A conscious human being is definitely an agent that is both capable of having intentions and capable of developing new ones. While the latter condition has been frequently implemented since Watt's governor, the latter one is still a rather unexplored feature of artificial agents. We could distinguish between teleologically fixed and teleologically open agents. A teleologically fixed agent is an agent whose goals (and possibly rules to define new ones) are fixed at design time. For instance, a robot whose goal is to pick up as many coke cans as it can is teleologically fixed – similarly, a software agent aiming at winning chess games or a controller trying to control the agent's movements in the smoothest possible way is teleologically fixed too. Learning usually does not affect goals. On the contrary, often learning is driven by existing goals.

On the contrary a human teenager is developing new kinds of goals all the time and most of them are only loosely connected with the phylogenetic instincts. Culture is a powerful example of phenomenon building on top of itself continuously swallowing new and unexpected goals. Is this a feature that is intrinsic of consciousness or it is only a contingent correlation between two otherwise separate phenomena?

It is a fact that artificial agents score very poorly in such area, 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 to develop a

new goal. For instance, in their seminal book on neural network learning processes, Richard S. Sutton and Andrew G. Barto stress 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 et al. 2006; Pfeifer et al. 2007), and their efforts were often related with machine consciousness.

It is interesting to stress the strong analogy between the two types of intentionality. Both are considered strongly correlated with conscious experience. Both are seen as intrinsic features of biological beings. Both play a role in linking mental states with external state of things. In short, it is fair to suspect that a conscious architecture could be the key to address both kind of intentionality and, indeed, it could be the case that the two types of intentionality are nothing but two aspects of consciousness.

20.2.6 Information Integration

Consciousness seems to be deeply related with the notion of unity. But what is unity in an artificial agent? The closest notion is that of information integration – namely, the necessity of unifying somewhere the many streams of otherwise separate and scattered flows of information coming from a multiplicity of heterogeneous sources.

In neuroscience, there is an analogous problem called the "binding problem" (Revonsuo and Newman 1999; Engel and Singer 2001; Bayne and Chalmers 2003). Such problem is usually called the "binding problem," and it has received no clear solution. Many binding mechanisms have been proposed ranging from temporal synchronization to hierarchical mechanism. So far, no one was satisfying.

Perhaps the most striking everyday demonstration of the binding problem is offered by our field of view. Watching at a normal scene with both eyes, we perceive a unified and continuous field of view with no gap between the left and the right side of it. And yet the visual processing is taking place in two separate areas, each in one of the two cerebral hemispheres. So to speak, if we watch someone walking from the right to the left of our field of view, the corresponding visual activities are going to shift from the left to the right hemisphere of our head. How can we explain the fact that we do not perceive any gap in the visual field, ever?

This notion is dangerously close to the easily discredited idea of an internal locus of control – an artificial version of the *homunculus* – a sort of Cartesian theatre where the information is finally presented to the light of consciousness. There are all sorts of objections to this view both from neuroscience and from AI. To start, possibly the view conflates together attention and consciousness, and this is not necessarily a good move as shown by all cases where the two takes place independently of one another (Koch and Tsuchiya 2006; Mole 2007; Kentridge et al. 2008).

So the question as to what gives unity to a collection of separate streams of information remains largely unanswered notwithstanding the everyday familiar experience of consciously perceiving the world a one big chunk of qualities. 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 the neuroscience wonders at the mystery of neural integration (Revonsuo 1999; Hurley 2003). 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 the necessity of a central locus of processing or the unity would 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 (1991) gain a unity eventually? Does a computer program have unity out of its programmer's head? Would embodiment and situatedness be helpful?

A possible and novel approach to this problem is the notion of integrated information introduced by 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 behavioral unity and autonomy of a conscious being is also associated with certain intrinsic causal and computational properties which could be responsible for having phenomenal experience (Tononi 2004).

20.2.7 Attention

If consciousness has to play a role in controlling the behavior of an agent, a mechanism which 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 it 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, 2007). Such model has several advantages since it suggests various ways to speed up the response and the accuracy of the agent.

The model is capable of making several predictions as to how to achieve better performances in dealing with incoming stimuli. Furthermore, it explains other features shared by the brain such as the competition between bottom-up signals from unexpected and strong inputs and top-down control aimed at goals. It also provides a mechanism to justify the inhibition between incoming signals. Other aspects addressed by the model are the Posner benefit effect in vision (Taylor and Fragopanagos 2005), some evolutionary and developmental aspects of the control systems, some features of the attentional blink, and some aspects of subliminal processing (Taylor and Fragopanagos 2007).

Another 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 and Rogers 2002; Taylor 2000, 2003, 2007).

It is worthwhile to describe briefly the structure of the CODAM model since it shows several common features with the model that will be described in the last part of this chapter. As shown in Fig. 20.1, the input enters through an incoming module generically labeled as input and is then sent both to a goal dedicated module and to an object map. In the object module is stored a set of higher level representation of

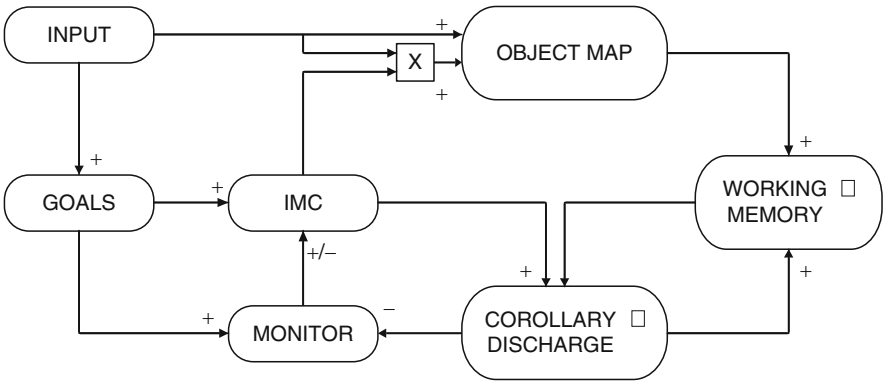

Fig. 20.1 The CODAM model architecture and the functional relations between its modules. Adapted from Taylor 2007, p. 987

previous relevant input objects. In this way, the incoming stimulus can be matched against the previously stored knowledge. At the same time, the input signal reaches the goal module, thus activating both an inverse module controller (IMC) and an internal monitor module. The role of the inverse controller is that of re-modulating the strength and thus the importance of the input signal as to the object module. On the other hand, the monitor module computes the difference between the desired goal state and the estimated state of the system.

20.3 A Consciousness-Oriented Architecture

In this paragraph, a tentative consciousness-oriented architecture is summarized. The architecture does not pretend to be either conclusive or experimentally satisfying. However, it is a cognitive architecture, so far only partially implemented (Manzotti 2003; Manzotti and Tagliasco 2005), whose design aims at coping with issues such as intentionality (of both kinds), phenomenal experience, autonomy, resilience, environment coupling, and information integration that are mandatory in order to address machine consciousness (Chella and Manzotti 2007, 2009). Or, at least, it ought to suggest how such issues are to be addressed. In many ways, it capitalizes on previous attempts and hopefully is making some predictions on what phenomenal content could be.

Before getting into the details of the suggested architecture, we would like to point out some other general principles that were used as guidelines in this work. These principles are derived both from theoretical considerations and from empirical facts about the brain.

- *Antirepresentationalist stance*. It is fair to say that the representationalist/antirepresentationalist debate has not been solved yet. Do conscious subjects perceive the world or their representations? Are internal representations the object of our mental life or are they just epistemic tools referring to slices of the causal chain that gets us acquainted with the environment? Should we prefer indirect or direct model of perception? In AI, representations were long held to be real. However, for many reasons outlined elsewhere (Manzotti 2006), a strong antirepresentationalist stance is here adopted: representations have a fictitious existence. Like centers of mass, they do not really exist. They are concepts with a high explanatory efficacy in describing the behavior of agents. Hence, in designing a conscious-oriented architecture, we better get rid of them and conceive our architecture as a part of the environment from the very start.
- *One architectural principle "to rule them all"*. The highly complex structure of the brain cannot be completely specified by the genetic code. Indeed, it could even be undesirable, since it would reduce neural adaptability. Further, many studies showed that most areas of the brain can emulate the neural circuitry of any other one (Sur et al. 1988; Roe et al. 1993; Sharma et al. 2000; Kahn and Krubitzer 2002; Hummel et al. 2004; Ptito et al. 2005). The blue print of the brain contained inside genes seems rather coarse. Probably, the highly complex final

structure of the brain, as it has been studied in neuroscience (Krubitzer 1995), is not the result of complex design specifications but rather of a few architectural principles capable of producing the final results (Aboitz et al. 2003). In this spirit, complex models are suspicious, simple rules able to generate complexity in response to environmental challenges are to be preferred.

– *Memory vs. speed*. It is not uncommon to utilize processor capable of Giga or even Teraflops. In terms of pure speed, the sheer brute force of artificial processors is no match for brain-ware. However, a somehow underestimated feature of the brain is the apparent identity between neural structures for memory and neural structures for processing. The rigid separation between the incredibly fast processor and the sequentially accessed memory of computers has no equivalent in the brain. Thus, it could make sense to consider architectures that would trade speed for some kind of memory integration, albeit implemented on Neumannesque hardware.
– *Embodied and situated cognition*. This is an issue strongly related with our antirepresentationalist stance. Representations are a trick devised to carry a replica of the world inside the agent. It is doubtful whether it is a successful strategy. Thus, from the start, the cognitive architecture will be taken as an element of a larger causal network made of both the body and of a relevant part of the environment too.
– *Phenomenal externalism*. This is, of course, the most provocative hypothesis – namely that the vehicles of phenomenal content are to be taken as the physical processes engaged between the agent and the environment and not as the properties of the symbolic processes taking place inside the agent. Bizarre as it may seem, the presented model has a twofold advantage. On one hand, it does not dwell either on any kind of dualism or on any ontologically diaphanous "inner reality." On the other hand, it makes predictions as to when phenomenal content occurs as a result of situated cognition. After all, the model suggest that phenomenal content is nothing but certain kind of physical processes.

The above mentioned are hunches that can be derived from various evidences. Yet they are not criteria to measure the success of such an architecture, whose goal is different from executing a task as in classic AI. A good check list could be provided by the topics considered in the previous paragraphs – i.e., intentionality of both kinds (Dennett 1987; Harnad 1995), phenomenal experience (Manzotti 2006), autonomy (Clark et al. 1999; Chella et al. 2001; Di Paolo and Iizuka 2008), resilience (Bongard et al. 2006), environment coupling (Robbins and Aydede 2009), and information integration (Tononi 2004; Manzotti 2009).

The architecture presented here is based on a causal structure that can be replicated again and again at different levels of complexity. The architecture will span three levels: the unit level, the module level, and the architecture level. A rather simple structural principle should be able to generate a complete architecture managing an unlimited amount of incoming information. The generated architecture is going to use all the memory at its disposal and it should not make an explicit distinction between data and processing.

20.3.1 The Elementary Intentional Unit

The basic element of our architecture is rather simple. It is a unit receiving an input (whether it be as simple as a bit or as complex as any data structure you could envisage) and producing an output (a scalar, an integer or a logical value). From many to one, so to speak. As we will see, the unit has also a control input, but this is a detail that will be specified below.

The main goal of the unit is getting matched with an arbitrary stimulus and, after such a matching, having the task of being causally related with such original stimulus. It could be implemented in many different ways. Formally, it can be expressed by an undefined function waiting its first input before being matched to it forever. Basically, it is like having a selective gate that is going to be burned on its first input. After being "burned," the unit has a significant output only if the current input resembles the first input. If a continuous similarity function and a continuous output are used, the gate tunes the passage of information rather than blocking/allowing it, yet the general principle remains the same.

The final detail is the control input signal. Due to the irreversible nature of the matching, it could make sense to have a way to signal when the first input is received. Since the unit could be activated only at a certain moment, it is useful preserving its potential until certain conditions are obtained. Thus the need for an external control signal that will activate the unit signaling that the first input is on its way.

Due to its role rather than to its elementary behavior, we label the unit as the *intentional unit*. It is important that, up to now, the unit is not committed to any particular kind of data or internal implementation. The unit is potentially very general. It can be adapted to any kind of inputs: characters, words, numbers, vectors, and images.

A more formal description will help getting the gist of the unit (Fig. 20.2). Any kind of input domain C can be defined. The output domain is more conveniently defined as a real number ranging from 0 to 1. Finally, a similarity function has to be

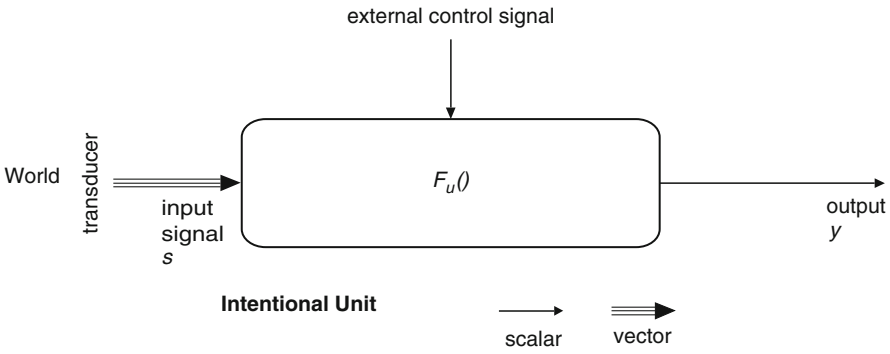

Fig. 20.2 The intentional unit

chosen – at worst, the identity function could be used. The similarity function $fs : C \times C \mathrm{a}[0, 1]$ must be such that standard conditions such as $fs(c, c) = 1, \forall c \in C$ & $fs(c_1, c_2) = fs(c_2, c_1), \forall c_1, c_2 \in C$ & $fs(c_1, c_2) \leq fs(c, c), \forall c_1, c_2, c \in C$ apply.

The similarity function is used to implement the intentional unit internal function $F_{\mathrm{u}} : C \mathrm{a}[0, 1]$ – the function that will change its behavior forever after its first input. F is defined as follows:

$$F_{\mathrm{u}}(s_t) = \begin{cases} 0 & t < t_0 \\ 1 & t = t_0 \\ fs_t(s_t, s_{t_0}) & t > t_0 \end{cases}.$$

It must be stressed that t_0 is an arbitrary chosen instant marked by the external control signal mentioned above. In short, the above function can be rewritten as:

$$F_{\mathrm{u}}(r, s_t) = v.$$

F_{u} is a function waiting for something to happen before adopting its final and fixed way of working. After the input $s_{t_0} \in C$ occurring simultaneously with the first time that $r = 1$, the output $y \in [0, 1]$ becomes the output of the similarity between the incoming input $s_t \in C$ and the original input s_{t_0}. The similarity function is as simple as the identity function or as complex as the designer likes. The most important requirement is that its maximum output value must hold for any value like the first received input.

Consider a few examples. Suppose that the incoming domain C is constituted by alphabetic characters, that the similarity function $fs : C \times C \mathrm{a}[0, 1]$ is the identity function, and that the output domain is the binary set $\{0, 1\}$. The function F_{u} is thus complete and the intentional unit can be implemented. The function F has no predictable behavior until it receives the first input. After that it will output 1 only when a character identical to the one received at its beginning is received. To conclude the example, imagine that a possible input is the following sequence of characters: "S," "T," "R," "E," "S," "S." The output will then be 1, 0, 0, 0, 1, 1.

A simple variation on the similarity function would permit a slightly fuzzier notion of similarity: two characters are similar (output equal to 1) either if they are the same or if they are next to each other in the alphabetical order. Given such a similarity function and the same input, the output would now be 1, 1, 1, 0, 1, 1. To make things more complex, a continuous output domain such as [0, 1] is admitted and the similarity function is changed to $fs\,(c) = 1 - (\mathrm{AD})/(\mathrm{TC})$, whereas AD is the alphabetical distance and TC is the total number of alphabetical character. With the above input, the output would then be: 1, 0.96, 0.96, 0.65, 1, 1. Clearly, everything depends on the first input. ,

A useful formalization of the unit is the one having vectors as its input domain. Given two vectors $\vec{v}, \vec{w} \in \Re^n$, a simple way to implement the similarity function is using a normalized version of a distance function between vectors $d\left(\vec{v}, \vec{w}\right)$, $d : (\Re^n \times \Re^n) \mapsto \Re$. Suitable candidates are the Minkowski function or the Tanimoto distance (Duda et al. 2001). Using grey images as input

vectors, elsewhere, a simple correlation function proved useful enough (Manzotti and Tagliasco 2005):

$$fs\left(\vec{v},\vec{w}\right)=\frac{1}{2}\left[1-C\left(\vec{v},\vec{w}\right)\right]=\frac{1}{2}\left[1-\frac{\sum\left(v_{\mathrm{i}}-\mu_{\mathrm{v}}\right)\left(w_{\mathrm{i}}-\mu_{\mathrm{w}}\right)}{\sqrt{\sum\left(v_{\mathrm{i}}-\mu_{\mathrm{v}}\right)^{2}\cdot\sum\left(w_{\mathrm{i}}-\mu_{\mathrm{w}}\right)^{2}}}\right].$$

These are all a few of the examples that could be made. It is important to stress that the unit is very adaptable and open to man different domains and implementation. Furthermore, the unit has a few features that are worth being stressed.

First, the unit does not distinguish between data and processing. In some sense, it is a unit of memory since its internal function is carved on a certain input thereby keeping a trace of it. Yet, there is no explicit memory, since there is not a stored value, but rather a variation in its behavior by means of its internal function. Indeed, the function can be implemented storing somewhere the value of the first input, but there is no need to have any explicit memory.

Another interesting aspect is that the unit shows a behavior which is the result of the coupling with the environment. When the unit is "burned," it is also forever causally and historically matched to a certain aspect of the environment (the one that produced that input).

Furthermore, there is no way to predict the unit future behavior since it is the result of the contingent interaction with the environment. If the input is unknown, the unit behavior is unknown too. If the input comes from the environment, there is no way to predict the unit behavior.

Finally, the unit seems to mirror, to a certain extent, some aspect of its own environment without having to replicate it. Slightly more philosophically, the unit allows to a pattern in the environment to exist by allowing it to produce effects through itself (more detailed considerations on this issue are outlined in Manzotti 2009).

By itself the intentional unit could seem pretty useless. However, things get more interesting once a large number of them are assembled together.

20.3.2 The Intentional Module

Suppose to have the capability of implementing and packing many intentional units into the same physical or logical package. The result could be a slightly more interesting structure here labeled as *intentional module*. This module has already been put to test in a very simplified robotic setup aiming at developing new motivations and controlling the gaze of a camera toward unexpected classes of visual stimuli (Manzotti and Tagliasco 2005; Manzotti 2007).

The simplest way to step from the intentional unit to the module is to pack together a huge number of intentional units all receiving the same input source. To avoid them behaving exactly the same, some mechanisms that prevents them from being shaped by the same input at the same time must be added. There are various ways to do it. A simple way is to number them and then to enforce that the units

can be burned sequentially. This could be obtained by means of the external control signal each intentional unit has. Because of its importance the external control signal is labeled *relevant signal*. The name expresses that such a signal is relevant in the life of each intentional unit since it controls to which input value the unit is matched forever.

Since the burning of each intentional unit will surely have a cost in terms of resources, it could make sense to add some more stringent conditions before switching on a relevant signal (and thus coupling forever an intentional unit to a certain input). Which conditions? To a certain extent they could be hard-wired and thus derived from some a priori knowledge the designer wants to inject into the system. Or, if the system is the result of some earlier generations of similar systems, it could be derived from the past history of previous generations. But it is definitely interesting whether the system could develop its own criteria to assign resources to further incoming inputs. By and large, the size of the input domain is surely much larger than that can be mapped by the available intentional units. In short, a feasible solution is having two explicitly divided sets of criteria working concurrently and then adding their outputs together so to have a relevant signal to be sent to the intentional units. The first set could be a set of hardwired functions trying to pin down external conditions that, for one reason or another, could be relevant for the system. The second set should be somehow derived from the growing set of matched intentional units themselves.

On the basis of what information these two sets of criteria should operate. A simple solution is on the basis of the incoming information with an important difference. The hardwired criteria could operate straight on the incoming data since they are hardwired and thus apply off-the-shelf rules. On the other hand, the derived set of criteria could use the output of the intentional units thereby using a historically selected subset of the incoming signals.

Shifting from the logical structure to the level of implementation, the above mentioned elements are packed into three submodules (Fig. 20.3). First the intentional units are packed into a huge array. Second, the hard-wired criteria are grouped together in such a way that they receive the signals jointly with the array of intentional units. Third, there is a third submodule grouping together the criteria derived by the past activity of the intentional units.

As it is clear from the structure, another small detail has been added: the module receives an external relevant signal that flanks the two internally generated ones. The reason for this will get clearer in the following paragraph.

A few more observations are needed. The module receives a vector (the incoming signal) and, possibly, an external relevant signal. It has two outputs: the internal relevant signal (the sum of hard-wired criteria, historically derived criteria, and possibly the external relevant signal) and a vector. For simplicity, all real values (both scalar signals and values of the vector elements) are normalized.

The output relevant signal is extremely important since it compresses all the past history of the system with respect to the current input signal.

The vector output is also a result both of the history and of the hard-wired criteria. Each element of this vector is the output of an intentional unit. Therefore, the output

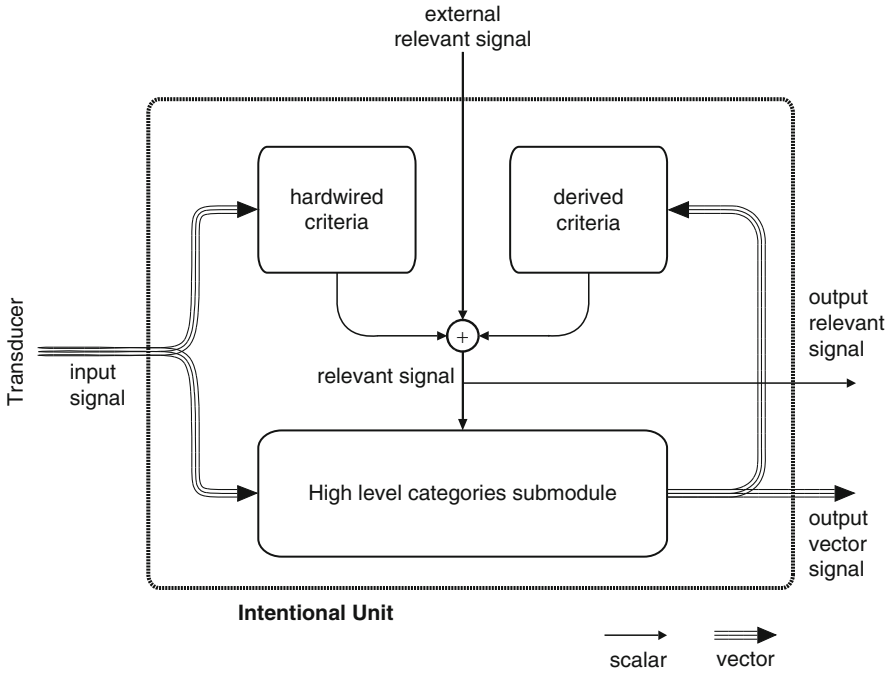

Fig. 20.3 The intentional module

vector has as many elements as there are intentional units. The value of each element expresses how much the corresponding intentional unit is activated by the current input signal, which in turn means how much the current input signal is similar to a given past input.

Formally, the intentional module implements a function:

$$F_{\mathrm{m}}\left(r, \vec{v}\right) = \begin{pmatrix} r_n \\ \vec{v}_n \end{pmatrix},$$

whereas r is the external control signal, $\vec{v}$ is the input vector, rn is the output relevant signal, and $\vec{v}_n$ is the output vector signal. Given an array of N intentional units, the output vector is:

$$\vec{v}_n = \begin{pmatrix} F_{\mathrm{u}}^{1}\left(r, \vec{v}\right) \\ \cdots \\ F_{\mathrm{u}}^{N}\left(r, \vec{v}\right) \end{pmatrix}.$$

It is interesting to note that each intentional unit (F_{u}^{i}) has a different starting time t_0^i, and thus it is matched to a different input vector $\vec{v}_i$.

20.3.3 The Intentional Architecture

The above intentional module is only slightly more interesting than the intentional unit, although it is already sufficient to implement classic conditioning, attentive behavior, and a rough self-generation of new goals (Manzotti 2009). There are a few features of the described intentional modules that are worth to be stressed once more.

First, the module can process any kind of data. It does not have to know in advance whether the incoming data is originated by images, sounds, texts, or whatever. In principle, at least, it is very general. Second, the module receives a vector and a scalar, and it outputs a vector and a scalar as well. There is ground to use the intentional module as the building block of a much larger architecture.

Second, the module uses vectors to send and receive data and scalars to send and receive controls.

Third, the module embeds its history in its structure. The module is unpredictable since its behavior is the result of a close coupling with the environment.

Now we will outline how to exploit these three features in order to design a more complex architecture.

Consider the case of a robot moving in an environment such as our own. In a real environment, there are multiple sources of information as well as multiple ways to extract different channels out of the same data source. Consider a visual color channel. It can be subdivided into a grey scale video, a color video, a filtered gray scale video (edges), and many other interesting channels. Besides, there are many more sources of information about the environment such as sound, tactile information, proprioception, and so on.

Consider having the capability of implementing a huge number of intentional modules. Consider having M incoming sources of information corresponding to as many vectors. For instance, vision could give rise to many different source of information. Sound capability could add a few more sources (different bandwidths, temporal vectors, spectral vectors), and so on. Suppose having M intentional modulestaking care of each of these sources. Suppose that each intentional module has a reasonably large number of intentional units inside. At this point, a few fixed rules will suffice to build dynamically an architecture made of intentional modules. Out of uniformity, we label it as *intentional architecture*. The rules are the followings:

1. Assign to each source of data a separate intentional module. Whether the capacity of the module is saturated (all intentional units are assigned), assign other intentional modules as needed. These modules make the first level of the architecture.
2. When the first level is complete, use the output of the first level modules as inputs for further levels of intentional modules.
3. The further level intentional modules are assigned to every possible earlier level modules output. However, due to many factors (the richness of the original external source of data, the implemented similarity function inside the intentional units, the incoming data, and so on), the output vector sizes are going to diminish

increasing the level. When this happens the intentional module will recruit a smaller and smaller number of intentional units. In that case, its output will get merged with that of other intentional modules with similarly smaller output.

4. In a while, the previous conditions will obtain for all intentional modules of the higher levels, thereby pushing toward a convergence.
5. All of the above applies for input and output vectors. As to the control signals, the rule is the opposite: backward connecting the higher level intentional modules with the lower level ones. In this way, the relevant signal produced by the highest possible intentional units will orient the activity of the lowest level intentional unit modules.
6. An example of a tentative final result is shown in Fig. 20.4. It has four sources of information (one of which is split onto two intentional units, thus providing five input signals). The end result is an architecture with four levels and twelve intentional modules (five modules in the first, four in the second, two in the third, and one in the fourth). Formally, the whole architecture behave like a giant module and its behavior could be expressed by the formula:

$$F_{\mathrm{a}}\left(r, \vec{v}_1 \ldots \vec{v}_M\right) = \begin{pmatrix} r_{\mathrm{a}} \\ \vec{v}_{\mathrm{a}} \end{pmatrix},$$

whereas r_{a} is the total relevant signal, $\vec{v}_a$ is the final output vector, and $\vec{v}_1 \ldots \vec{v}_M$ is a series of input signals. And r? Although it is not represented in Fig. 20.4, it cannot be excluded that that the architecture would admit a global relevant signal for all its modules. If this were the case, r_{a} would be the global relevant signal alerting all of the modules that something relevant is indeed arriving.

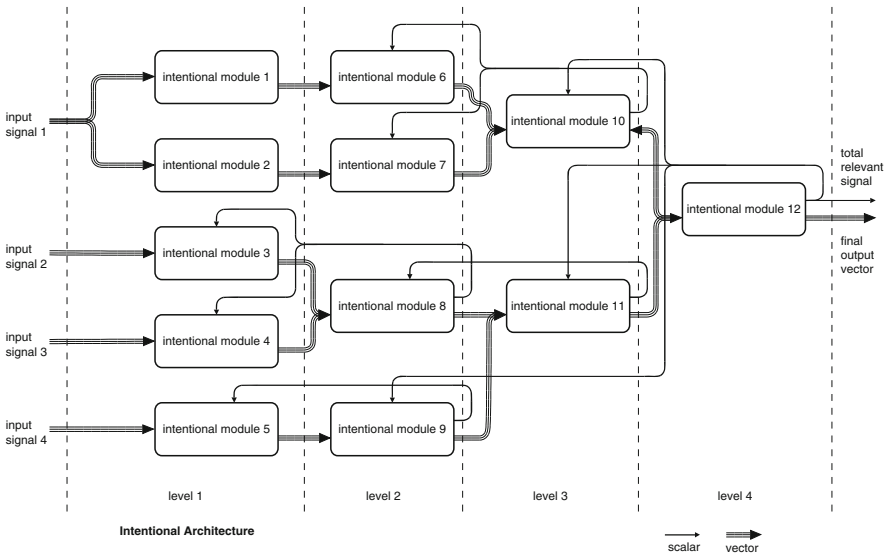

Fig. 20.4 The intentional architecture

20.3.4 Check List for Consciousness-Oriented Architectures

It is time to use the checklist for consciousness-oriented architectures suggested above. Consider the suggested intentional architecture as a test bed to check the issues mentioned above.

Bottom-up vs. top-down. The architecture is neither bottom-up nor top-down. On one hand, it is built from scratch by the incoming data. On the other hand, at every level it sends backward the relevant control signal thereby tuning the way in which the information recruits new intentional units and new intentional modules. In some sense, the final choice depends on the highest level relevant signal. However, such a signal is the result of all the bottom-up activity taking place before.

Information integration. All the incoming information flows through the system producing two final outputs (a vector and a scalar) that, in different ways, summarize all that happened before. It could be interesting to apply Tononi's measures to the information flow of it. Further, the integration happens at every step (both at the unit level, at the module level, and at the architecture level). The existence of backward control signals ensures that the structure is not a simple feed-forward one, but rather a complex recurrent one.

Environment coupling. All the information received in the past gets embedded into the structure of the architecture. The history of the architecture is embedded into its structure. Each intentional units is matched to a past input signal and thus to a past event. The architecture is carved out by the environment step by step and, in a deep causal sense, it mirrors the structure of the world to a certain extent.

Intentionality of the first kind. Here deep philosophical ground has to be addressed. Intentionality as semantics is not going be an easy game. Yet, the proposed architecture suggests some useful approach. The external events, which are the source of the input, are not causally inactive (for the architecture, at least) before being matched whether to an intentional unit or to a cluster of intentional units. After the matching, the event has a new causal role. Consider a pattern. Before it being seen by someone, does it really exist? Similarly a pattern in the world gets coupled with the activity inside the architecture by means of the causal relation endorsed by the intentional unit. Such a causal relation is the same one that allows that very pattern to be causally active. The outlined causal structure could be a way to tackle with the problem of semantics. It could be the way to naturalize it.

Intentionality of the second kind. The system has both hardwired fixed criteria and newly developed criteria. Each intentional module can develop new criteria that depend only on its past history. In a similar way, the architecture is combining all the relevant signals into a global relevant signal that could be the foundation for a global motivational system. It is a teleologically open system.

One architectural principle to rule them all. This requirement has been satisfied rather well. The architecture is the repetition of the same causal structure again and again. From the intentional unit up to the architecture itself, the causal struc-

ture is always following the same dictum: receive inputs and change accordingly as to become causally sensitive to those very inputs. Match yourself with the environment! This is rather evident if we compare the three functions describing the behavior of the unit level, the module level, and the architecture level:

$$F_{\mathrm{u}}(r, s_{\mathrm{t}}) = v,$$

$$F_m\left(r, \vec{v}\right) = \begin{pmatrix} r_n \\ \vec{v}_n \end{pmatrix},$$

$$F_{\mathrm{a}}\left(r, \vec{v}_1 \ldots \vec{v}_M\right) = \begin{pmatrix} r_{\mathrm{a}} \\ \vec{v}_{\mathrm{a}} \end{pmatrix}.$$

The only significant difference is the increase in the complexity of the incoming and outgoing information. It will not be difficult to envisage an even higher level made of whole architectures combining together.

Antirepresentationalist stance. There are neither explicit representations nor variable stored anywhere. Every complex incoming cluster of events is distributed in all the architecture as a whole. There is no way to extract a particular representation from the values stored inside the architecture. However, when exposed to a certain stimulus, the architecture will recognize it. The architecture has a meaning only if considered as a whole with the environment. It is both embodied and situated.

Autonomy and resilience. First, it is interesting to note that the architecture could suffer substantial damage both in the intentional units and even in the intentional modules without being totally destroyed. For instance, the loss of a whole module would imply the loss of the capability to deal with the corresponding information source. Yet, the system will continue to behave normally with respect to all other information sources. As to the other sources, the system will show no sign of reduced performances. Similarly, new information sources could be added anytime, although if they are present at the very beginning, the architecture will incorporate them more seamlessly. As a result, the architecture resilience seems pretty good. As to the autonomy, it can be stressed that the system develops both epistemic categories (by means of intentional units of higher levels) and goals (by means of the backward control signals). In this way, the system is going to be rather unpredictable and strongly coupled with its environment thereby implementing a strong decisional and epistemic autonomy.

Phenomenal experience. It would be unfair to pretend a final proof of the occurrence of phenomenal experience since its final solution will probably require some ontological breakthrough whose scope is definitely a lot more far reaching than cognitive science and artificial intelligence could envisage. Yet, the present model is flanked by the externalist view that so far has neither being validated nor rejected. We think it is possible to foresight that this kind of architecture (or a much improved one in the same spirit) would start to address phenomenal experience since the kind of engagement between such an architecture and the environment could lead to the continuity required to have phenomenal experience. Consider that all the

events occurring inside the architecture are the result of environmental phenomena and that the emergence of top-down signal is the eventual outcome of stimuli originating externally to the architecture. The architecture is significantly shaped by the environment.

20.3.5 A Comparison with Other Approaches

It is useful to compare other attempts at modeling a conscious machine. The following is by no means neither an exhaustive list nor a throughout description of the architectures mentioned. However, for those already aware of other approaches a quick glance could help to gain a better grasp of the gist of the architecture presented here. One more caveat is that we will focus on differences rather than on the commonalities.

The most obvious candidate is Taylor's CODAM model of attention. We would like to point out at the similarity between the intentional module and the CODAM model previously quoted and sketched. Compare the two diagrams in Figs. 20.1 and 20.3. In both cases, a similar interplay between control and memory occurs. Both models have separate and parallel control signals going from the input either to the memory or to the controller. A few models, and their connections too, can be rather convincingly matched against each other. The CODAM object module is similar in role and in internal structure to the category array of the intentional module. Both store high level representations of the stimuli received. Both have a control signal which is the result of the combination of a goal system and the bottom-up contributions of the match between the incoming stimuli and the learning achieved. In the CODAM model, this is achieved by means of the cooperation between the goal module, the IMC, and the working memory. In the presented intentional module, this is achieved thanks to the sum of two signals: the output of the hardwired criteria submodule, which is analogous to the goal module, and the output of the acquired criteria submodule, which receives from the category submodule (as the working memory too is fed by the object map in the CODAM model).

Of course, there are also many relevant differences. The CODAM model is closer to the brain neural architecture, and it is particularly suited to match many attentional processes in the visual system. On the other hand, the intentional architecture does not try to explain either attentional or the cognitive features of biological systems, rather it capitalizes on them in order to achieve information integration, environmental coupling, autonomy, and intentionality. Whether it is successful, it is matter of further experimental research.

Another candidate for comparison 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, 2003). They aim at a full functional integration between competing software agencies. The biggest difference with our approach is that we make explicit hypothesis as to how information has to

be processed in order to reproduce the same properties of a conscious brain. While Franklin's approach is fully compatible with the functionalistic tenet of multiple realizability, our approach isn't. 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 Baars' Global Workspace as it has been implemented by 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 and, 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 that mixes together the more classic cognitive structure of global workspace with largely not symbolic neural networks.

Both Shanahan's and our approach exploit a tight coupling between the environment and the agent. Yet, our architecture does not aim at modeling explicitly either a common workspace or a predesigned organization between various cognitive functions. Rather, the "conscious" bottleneck, so to speak, ought to stem out of the bottom-up vs. top-down interplay described above. Another important difference is that, at least conceptually, Shanahan's approach assumes that there is a separation between the inside and the outside of the agent. Inside the agent, it is meaningful to speak of information, computation, broadcasting, and representations. Outside, there are just physical events. Shanahan's architecture matches such theoretical chasm. On the contrary, since we stress the concrete possibility that consciousness could be spread in the environment, our architecture is designed as to become seamlessly causally connected with its historical and individual surroundings.

Similar considerations hold for Hesslow's emphasis on emulation (Hesslow and Jirenhed 2007). Since he is apparently committed to the view that consciousness inheres on the existence of an inner world, Hesslow defends the so-called simulation hypothesis which does not rely on any assumptions about the nature of imagery or perception except that activity in sensory cortex can be elicited internally. He conceives emulation as an internally generated feedback loop. On the contrary, in the presented architecture, everything can always be traced back to something occurred externally. To a certain extent, our architecture could be criticized for being almost completely environment-driven. Yet this is not a shortcoming but rather an unavoidable condition shared by agents once they are seen as a part of the environment rather than as a separate domain.

As for the broad category of approaches explicitly referring to embodiment and situatedness, we can quote both Holland's work on *Cronos* and Bongard on resilient

autonomous agents (Holland 2004; Bongard et al. 2006; Pfeifer and Bongard 2006; Pfeifer et al. 2007). Both cases highlight the importance of embodiment. However, although they suggest the importance of taking advantage of morphology and embodiment to achieve superior performance, they still seem to consider that models and representations are instantiated inside the cognitive domain of the agent. So the world acts either as an external memory or as an external collection of functional modules or, finally, as a simplifier of otherwise too complex computational problems. We do pursue a different approach. Our architecture tries to exploit the causal structure of the environment to develop accordingly. However, it remains incomplete whether not coupled with the causal history it triggered its growth. Yet, the spirit of our architecture is definitely closer to such embodied approach rather than to the pure functionalist approach of classic AI.

There are many more authors we could only quote such as Chrisley's synthetic phenomenology, Aleksander's axioms, Haikonen's conscious machines, and many others (Haikonen 2003; Aleksander et al. 2008; Aleksander and Morton 2008; Chrisley 2009). Yet it is perhaps more useful to outline two critical dimensions along which most attempts lie. The first dimension is the dichotomy between internalist approaches and externalist approaches passing through embodiment and situatedness. The second dimension corresponds to the dichotomy between pure multiple realizability vs. critical style of information processing. The former view assumes that as long as two agents perform to share the same functional behavior, they are the same. The latter view assumes that, at some level, it matters the way in which a certain cognitive task is performed. We can thus sketch in Fig. 20.5 a recapping map of the various views and available architectures.

A final remark. Most of the aforementioned differences are rooted in the different conceptual frameworks adopted by authors. Thus, it is possible that, once expressed using a completely neutral jargon, two randomly picked up architectures would end up being much closer than they would appear when disguised under their authors' theoretical commitments. Yet this is not enough to set aside all differences. At this stage of research, an architecture is not simply either a functional or a mathematical diagram. An architecture is a bundle with the viewpoint it endorses. To make an example from the recent history of AI, consider neural networks and not linear mathematical approximators. At a certain descriptive levels, they are the same, but if represented in their theoretical framework they look different. Neural networks represented a cognitive tool whose efficacy went much beyond its mathematical properties for better or for worse.

20.4 Conclusion

Although AI achieved impressive results (Russell and Norvig 2003), it is always astonishing the degree of overvaluation that many nonexperts 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

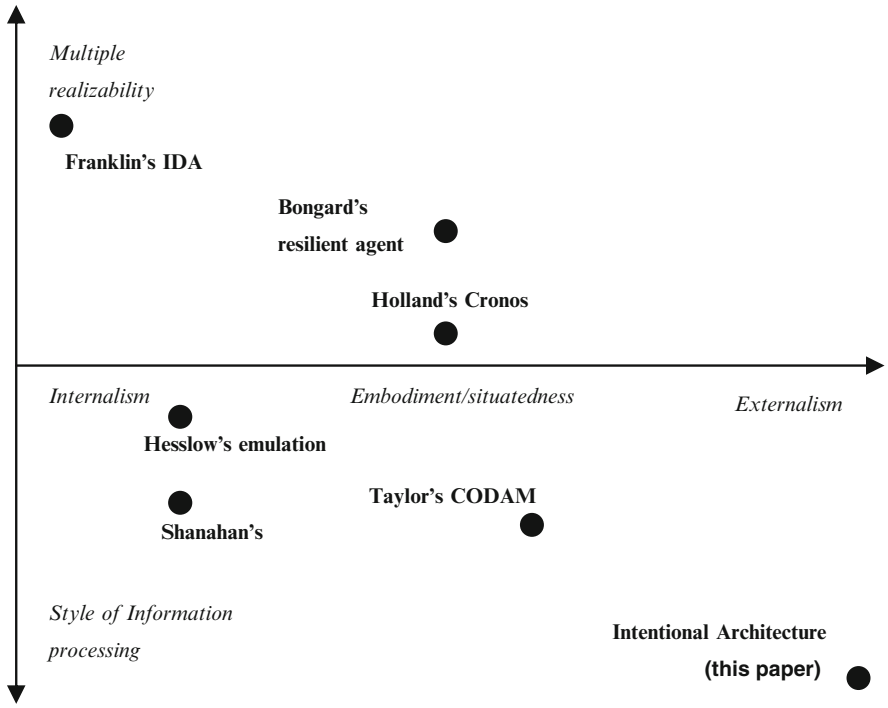

Fig. 20.5 A map of current attempt at achieving machine consciousness

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 current status of AI research. Occasionally, even skilled scholars slip into this habit quoting implausible Legobots used in first-year undergraduate robot instruction as agents capable of developing new motivations (Aleksander et al. 2008, pp. 102–103).

Such approximate misevaluations of the real status of AI hinder 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.

Studying consciousness and suggesting artificial model of it could help addressing those aspects of the mint that have remained so far elusive. The available results encourage to keep designing conscious machines or, at least, machines that reproduce some aspects of conscious experience. Although a complete theory of consciousness will very probably require some ground-breaking conceptual revisions, the deliberate attempt at mimicking consciousness will pave the way to a future scientific paradigm to understand the mind.

References

Aboitz, F., D. Morales, et al., (2003), "The evolutionary origin of the mammalian isocortex: Towards an integrated developmental and functional approach." in *Behavioral and Brain Sciences*, 26: 535–586.

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

Adams, D., K. Aizawa, (2008), *The Bounds of Cognition*, Singapore, Blackwell.

Adams, F., K. Aizawa, (2009), "Why the Mind is Still in the Head" in P. Robbins and M. Aydede, Eds, *The Cambridge Handbook of Situated Cognition*, Cambridge, Cambridge University Press: 78–95.

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

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

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

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

Aleksander, I., H. Morton, (2008), "Depictive Architectures for Synthetic Phenomenology" in A. Chella and R. Manzotti, Eds, *Artificial Consciousness*, Exeter, Imprint Academic(30–45).

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

Atkinson, A. P., M. S. C. Thomas, et al., (2000), "Consciousness: mapping the theoretical landscape." in *TRENDS in Cognitive Sciences*, 4(10): 372–382.

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

Bayne, T., D. Chalmers, (2003), "What is the Unity of Consciousness?" in A. Cleeremans, Ed., *The Unity of Consciousness: Binding, Integration, and Dissociation*, Oxford, Oxford University Press: 23–58.

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

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

Brentano, F., (1874/1973), *Psychology From an Empirical Standpoint*, London, Routledge & Kegan Paul.

Bringsjord, S., (1994), "Computation, among other things, is beneath us." in *Minds and Machines*, 4(4): 469–488.

Brooks, R. A., (1990), "Elephants Don't Play Chess." in *Robotics and Autonomous Systems*, 6: 3–15.

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

Brooks, R. A., C. Breazeal, et al., (1998), "Alternate Essences of Intelligence", in *AAAI 98*.

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(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.

Changeux, J. P., (2004), "Clarifying Consciousness." in *Nature*, 428: 603–604.

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., R. Manzotti, Eds, (2007), *Artificial Consciousness*, Exeter (UK), Imprint Academic.
Chella, A., R. Manzotti, (2009), "Machine Consciousness: A Manifesto for Robotics." in *International Journal of Machine Consciousness*, 1(1): 33–51.
Chrisley, R., (1994), "Why Everything Doesn't Realize Every Computation." in *Minds and Machines*, 4(4): 403–420.
Chrisley, R., (2008), "The philosophical foundations of Artificial Consciousness." in *Journal of Artificial Intelligence and Medicine*(Special Issue on Artificial Consciousness).
Chrisley, R., (2009), "Synthetic Phenomenology." in *International Journal of Machine Consciousness*, 1(1): 53–70.
Clark, A., (1997), Being there: putting brain, body and world together again, Cambridge (Mass), MIT.
Clark, A., (2008), *Supersizing the Mind*, Oxford, Oxford University Press.
Clark, A., D. Chalmers, (1999), "The Extended Mind." in *Analysis*, 58(1): 10–23.
Clark, O. G., R. Kok, et al., (1999), "Mind and autonomy in engineered biosystems." in *Engineering Applications of Artificial Intelligence*, 12: 389–399.
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.
Crick, F., (1994), *The Astonishing Hypothesis: the Scientific Search for the Soul*, New York, Touchstone.
Crick, F., C. Koch, (1990), "Toward a Neurobiological Theory of Consciousness." in *Seminars in Neuroscience*, 2: 263–295.
Crick, F., C. Koch, (2003), "A framework for consciousness." in *Nature Neuroscience*, 6(2): 119–126.
Dennett, D. C., (1978), *Brainstorms: philosophical essays on mind and psychology*, Montgomery, Bradford Books.
Dennett, D. C., (1987), *The intentional stance*, Cambridge (Mass), MIT.
Dennett, D. C., (1991), *Consciousness explained*, Boston, Little Brown and Co.
Di Paolo, E. A., H. Iizuka, (2008), "How (not) to model autonomous behaviour." in *BioSystems*, 91: 409–423.
Dretske, F., (1985), "Machines and the Mental." in *Proceedings and Addresses of the American Philosophical Association* 59(1): 23–33.
Duda, R. O., P. E. Hart, et al., (2001), *Pattern classification*, New York, Wiley.
Engel, A. K., W. Singer, (2001), "Temporal binding and the neural correlates of sensory awareness." in *TRENDS in Cognitive Sciences*, 5: 16–25.
Franklin, S., (1995), *Artificial Minds*, Cambridge (Mass), MIT.
Franklin, S., (2003), "IDA: A Conscious Artefact?" in *Journal of Consciousness Studies*, 10: 47–66.
Gallagher, S., (2009), "Philosophical Antecedents of Situated Cognition" in P. Robbins and M. Aydede, Eds, *The Cambridge Handbook of Situated Cognition*, Cambridge, Cambridge University Press.
Haikonen, P. O., (2003), *The Cognitive Approach to Conscious Machine*, London, Imprint Academic.
Hameroff, S. R., A. W. Kaszniak, et al., (1996), *Toward a science of consciousness: the first Tucson discussions and debates*, Cambridge (Mass), MIT.
Harnad, S., (1990), "The Symbol Grounding Problem." in *Physica*, D(42): 335–346.
Harnad, S., (1994), "Computation is just interpretable symbol manipulation; cognition isn't " in *Minds and Machines*, 4(4): 379–390.
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., 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).

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.

Hesslow, G., D.-A. Jirenhed, (2007), "Must Machines be Zombies? Internal Simulation as a Mechanism for Machine Consciousness", in *AAAI Symposium*, Washington DC, 8–11 November 2007.

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.

Hummel, F., C. Gerloff, et al., (2004), "Cross-modal plasticity and deafferentiation." in *Cognitive Processes*, 5: 152–158.

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.

Hurley, S. L., (2006), "Varieties of externalism" in R. Menary, Ed., *The extended mind*, Aldershot, Ashgate publishing.

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

Kahn, D. M., L. Krubitzer, (2002), "Massive cross-modal cortical plasticity and the emergence of a new cortical area in developmentally blind mammals." in *Proceedings of National Academy of Science*, 99(17): 11429–11434.

Kentridge, R. W., T. C. W. Nijboer, et al., (2008), "Attended but unseen: Visual attention is not sufficient for visual awareness." in *Neuropsychologia*, 46: 864–869.

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

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

Koch, C., N. Tsuchiya, (2006), "Attention and consciousness: two distinct brain processes." in *TRENDS in Cognitive Sciences*, 11(1): 16–22.

Krubitzer, L., (1995), "The organization of neocortex in mammals: are species differences really so different?" in *Trends in Neurosciences*, 18: 408–417.

Libet, B., C. A. Gleason, et al., (1983), "Time of conscious intention to act in relation to onset of cerebral activity. The unconscious initiation of a freely voluntary act." in *Brain*, 106(3): 623–642.

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

Mack, A., I. Rock, (1998), *Inattentional Blindness*, Cambridge (Mass), MIT.

Manzotti, R., (2003), "A process based architecture for an artificial conscious being" in J. Seibt, Ed., *Process Theories: Crossdisciplinary studies in dynamic categories*, Dordrecht, Kluwer: 285–312.

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

Manzotti, R., (2006), "Consciousness and existence as a process." in *Mind and Matter*, 4(1): 7–43.

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

Manzotti, R., (2009), "No Time, No Wholes: A Temporal and Causal-Oriented Approach to the Ontology of Wholes." in *Axiomathes*, 19: 193–214.

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

Manzotti, R., V. Tagliasco, (2008), "Artificial Consciousness: A Discipline Between Technological and Theoretical Obstacles." in *Artificial Intelligence in Medicine*, 44(2): 105–118.

McFarland, D., T. Bosser, (1993), *Intelligent Behavior in Animals and Robots*, Cambridge (Mass), MIT.

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

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

Miller, G., (2005), "What is the Biological Basis of Consciousness?" in *Science*, 309: 79.
Mole, C., (2007), "Attention in the absence of consciousness." in *TRENDS in Cognitive Sciences*, 12(2): 44–45.
Nemes, T., (1962), *Kibernetic Gépek*, Budapest, Akadémiai Kiadò.
Nemes, T., (1969), *Cybernetic machines*, Budapest, Iliffe Books and Akademiai Kiadò.
Newell, A., (1990), *Unified Theories of Cognition*, Cambridge (Mass), Harvard University Press.
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., 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., (2007), "Self-Organization, Embodiment, and Biologically Inspired Robotics." in *Science*, 5853(318): 1088 - 1093.
Pockett, S., (2004), "Does consciousness cause behaviour?" in *Journal of Consciousness Studies*, 11(2): 23–40.
Prinz, J., (2000), "The Ins and Outs of Consciousness." in *Brain and Mind*, 1: 245–256.
Prinz, J., (2009), "Is Consciousness Embodied?" in P. Robbins and M. Aydede, Eds, *The Cambridge Handbook of Situated Cognition*, Cambridge, Cambridge University Press: 419–436.
Ptito, M., S. M. Moesgaard, et al., (2005), "Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind." in *Brain*, 128: 606–614.
Revonsuo, A., (1999), "Binding and the phenomenal unity of consciousness." in *Consciousness and Cognition*, 8: 173–85.
Revonsuo, A., J. Newman, (1999), "Binding and consciousness." in *Consciousness and Cognition*, 8: 127–127.
Robbins, P., M. Aydede, Eds, (2009), *The Cambridge Handbook of Situated Cognition*, Cambridge, Cambridge University Press.
Rockwell, T., (2005), *Neither ghost nor brain*, Cambridge (Mass), MIT.
Roe, A. W., P. E. Garraghty, et al., (1993), "Experimentally induced visual projections to the auditory thalamus in ferrets: Evidence for a W cell pathway." in *Journal of Computational Neuroscience*, 334: 263–280.
Rowlands, M., (2003), *Externalism. Putting Mind and World Back Together Again*, Chesham, Acumen Publishing Limited.
Russell, S., 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., (1980), "Minds, Brains, and Programs." in *Behavioral and Brain Sciences*, 1: 417–424.
Searle, J. R., (1992), *The Rediscovery of the Mind*, Cambridge (Mass), MIT.
Seth, A., (2009), "The Strength of Weak Artificial Consciousness." in *International Journal of Machine Consciousness*, 1(1): 71–82.
Shanahan, M., (2006), "A Cognitive Architecture that Combines Internal Simulation with a Global Workspace." in *Consciousness and Cognition*, 15: 433–449.
Shanahan, M., B. J. Baars, (2005), "Applying Global Workspace Theory to the Frame Problem." in *Cognition*, 98(2): 157–176.
Shanahan, M. P., (2005), "Global Access, Embodiment, and the Conscious Subject." in *Journal of Consciousness Studies*, 12(12): 46–66.
Sharma, J., A. Angelucci, et al., (2000), "Visual behaviour mediated by retinal projections directed to the auditory pathway." in *Nature*, 303: 841–847.
Simons, D. J., (2000), "Attentional capture and inattentional blindness." in *TRENDS in Cognitive Sciences*, 4: 147–155.
Simons, P. M., (1987), *Parts. A Study in Ontology*, Oxford, Clarendon.

Skrbina, D., Ed. (2009), *Mind that abides. Panpsychism in the new millennium*, Amsterdam, John Benjamins Pub.

Sur, M., P. E. Garraghty, et al., (1988), "Experimentally induced visual projections into auditory thalamus and cortex." in *Science*, 242: 1437–1441.

Sutton, R. S. and A. G. Barto, (1998), *Reinforcement Learning*, Cambridge (Mass), MIT.

Tagliasco, V., (2007), "Artificial Consciousness. A Technological Discipline" in A. Chella and R. Manzotti, Eds, *Artificial Consciousness*, Exeter, Imprint Academic: 12–24.

Taylor, J. G., (2000), "Attentional movement: The control basis for consciousness." in *Society for Neuroscience Abstracts*, 26: 2231#839.3.

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

Taylor, J. G., (2003), "Neural Models of Consciousness" in M. A. Arbib, Ed., *The Handbook of Brain Theory and Neural Networks*, Cambridge (Mass), MIT: 263–267.

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.

Taylor, J. G. and N. Fragopanagos, (2005), "The interaction of attention and emotion." in *Neural Networks*, 18(4): 353–369.

Taylor, J. G. and N. Fragopanagos, (2007), "Resolving some confusions over attention and consciousness." in *Neural Networks*, 20(9): 993–1003.

Taylor, J. G. and M. Rogers, (2002), "A control model of the movement of attention." in *Neural Networks*, 15: 309–326.

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

Van Gelder, T., (1995), "What Might Cognition Be, If Not Computation?" in *The Journal of Philosophy*, 92(7): 345–381.

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

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

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.