From the Perspective of Artificial Intelligence: A New Approach to the Nature of Consciousness

Riccardo Manzotti
Department of Linguistics and Philosophy
Massachusetts Institute of Technology
MA, United States

Sabina Jeschke
Institute Cluster IMA/ZLW & IfU
RWTH Aachen University
Aachen, Germany

***Abstract*—Consciousness is not only a philosophical but also a technological issue, since a conscious agent has evolutionary advantages. Thus, to replicate a biological level of intelligence in a machine, concepts of machine consciousness have to be considered. The widespread internalistic assumption that humans do not experience the world as it is, but through an internal '3D virtual reality model', hinders this construction.**

To overcome this obstacle for machine consciousness a new theoretical approach to consciousness is sketched between internalism and externalism to address the gap between experience and physical world. The 'internal interpreter concept' is replaced by a 'key-lock approach'. Here, consciousness is not an image of the external world but the world itself.

A possible technological design for a conscious machine is drafted taking advantage of an architecture exploiting self-development of new goals, intrinsic motivation, and situated cognition. The proposed cognitive architecture does not pretend to be conclusive or experimentally satisfying but rather forms the theoretical the first step to a full architecture model on which the authors currently work on, which will enable conscious agents e.g. for robotics or software applications.

Keywords—consciousness; machine consciousness; multi agent system; genetic algorithms; externalism

I. Introdution

Even if consciousness is not exactly a 'well-defined' term and generations of philosophers and other scientists have discussed its complex features at length, there is a certain common understanding about its central meaning: Consciousness describes the unique capability of having experiences in terms of perceptions, thoughts, feelings and awareness.1 Obviously, consciousness requires the awareness of the external world. What is still fairly mysterious is the nature of this experience. Although this capability is still very poorly understood and indeed is considered a sort of challenge for the standard picture of the world, it is a plain fact that the conscious human being is one of the outcomes of natural selection. Likewise, it seems undeniable that human beings cope with the most unexpected events by means of conscious reflection. Finally, they are extremely sensitive to anything remotely resembling the capability of feeling in other agents. In sum, consciousness appears to be a not negotiable aspect of a highly developed autonomous agents and it cannot be underestimated that the practical advantages may result from its replication within an artificial being [1] [2] [3] [4] [5]. Here, the problem of the physical underpinnings of consciousness rather than the problem of the self is addressed and thus the nature of what it is like to have a certain experience [6] [7] [8] rather than the problem of how the different cognitive processes combine together to form a self. This paper considers how experience may be the result of a physical system interaction with the world – experience rather than the self is the goal of this proposal.

During the recent decades, one got familiar with the conception that one never gets acquainted to the world as it is, but only to a '3D virtual reality model' of the outside world that one's brain switches on as soon as one wakes up. This internal model is taken to be the inner world of consciousness – how the world appears to humans and not what the world really is. To give an example, colour in the external world may be defined, albeit with some simplification, by two physical parameters: wavelength and intensity. For a human being however, light is not just the detection of a certain light frequency on the retina, but a certain experience when detecting that light frequency (say, perceiving red). In his excellent textbook on vision, Stephen Palmer claims: 'Color is a psychological property of our visual experiences when we look at objects and lights, not a physical property of those objects or lights' [9]. Nevertheless, this psychological property is without comprehensive explanation so far. It does not fit to any obvious physical property.

In a nutshell, the current main line of explanation goes as follows:

- An external event
- goes through some kind of internal interpreter (in the human brain)
- and internally produces a certain result (within the human).

This current interpretation can be allocated to the so-called 'internalistic models' – namely those models that take the mind to be a property of what takes place inside the neural system

[1] Consciousness is an 'umbrella term' encompassing a variety of distinct meanings. For this purpose it is important to differentiate between consciousness and self-consciousness. In this paper, to be conscious it is only necessary to be aware of the external world, whereas self-consciousness is an acute sense of self-awareness. In Chapter IV.A. this topic is discussed in more detail.

[10] [11] [12][2]. This view has its strength: putting some kind of 'interpretation layer' between the individual and the outside world allows explaining why humans – all of them build alike – tend to be rather different in their behaviours, reactions and 'feelings'. This argument can be extended to non-human species as well: if the same physical reality 'shows itself' differently to diverse entities, a tentative explanation for the heterogeneous behaviours of these unequal species living in the same environment may be put forward. It also allows an explanation of why humans seem to be capable of consciousness in the absence of obvious external stimuli – in the case of dreams, hallucinations, and afterimages.

However, there are still many open ends, some riddles concerning the conception of consciousness that cannot be answered with the abovementioned picture:

- First, to believe that 'the internal interpreter', and 'neurons will do it "somehow" ', is interesting but doubtful. In the last couple of decades, scientists from all areas have invested a lot of energy in the quest for a neural mechanism capable of producing our everyday conscious experience. Up to now, there is no known law of nature predicting that neural activity should result in one's experience.
- Second, in the current model, consciousness neither fits the physical world nor its properties. To carry it to the extremes, that means that one constantly ignores the 'real world' by overwriting it with some internal 'fantasies'. Of course, this could be the case – but it sounds at least pretty counterintuitive: why should nature take this kind of detour?
- Third: The discrepancy between our immediate experience and the 'world' is more than just 'somewhat regrettable'. If everything one experiences – from pain to colour, from pictures to music – is nothing more than a product of human neurons, then a logical problem occurs: Why should it be easier for neurons to transmute neural firings into music – than for a cello to shape airwaves into music? If the physical world is devoid of qualitative features, why should the brain – which is part of the physical world – be any better in this respect? Why should the brain create meaningful things, but a cello does not? Or, to use an even catchier picture: 'If colours cannot pop out of strawberries, how can they pop out of neurons?'

As anticipated at the beginning of this section, consciousness is not only a scientific conundrum but a practical goal, too. From the Artificial Intelligence community (and the authors admit that they belong to that community), another thought comes up: Whatever consciousness is in detail, it seems to form an important part at least of a human-like intelligence [13] [14] [15] [16] [17] [18] [19]. Therefore, to build artificial systems with certain intelligence, it might be necessary to give them some kind to consciousness too – even if this artificial consciousness might differ very much from the human one, or from other mammals or biological systems. Now, however consciousness might work in biological systems, one may envisage implementing consciousness in totally different ways as part of forthcoming technological systems. Therefore, alternative models to explain consciousness are of the utmost interest, either to explain the 'true nature of human consciousness', or, to allow different approaches for building an artificial/technological agent.

This paper proposes a new hypothesis concerning the nature of conscious experience, to overcome a conceptual war between externalism and internalism (Chapter II). In Chapter III the consequences of this new perspective are discussed. Chapter IV relates the change of perspective to the field of artificial intelligence. Finally, Chapter V summarizes the paper and gives an outlook on the next research steps to be undertaken.

II. Towards a New Conceptualization of Consciousness

A. The approach

This chapter's goal is to flesh out a new concept of consciousness that is directed to overcome the gap between 'experience' and 'physical world'. To reach that goal, one has to undergo several changes in the standard mindset. Below the hardest nuts to crack are mentioned:

1) To bridge the gap between experience and physical world, the external surrounding environment in which a brain and a body are situated have to be more prominent in our concept of consciousness (since the other way around seems even more radical).

2) If the 'externality' becomes more important, the next domino falling is the giving-up of the underlying 'full-bodied' human-centered view of being necessarily located fully and totally inside its body. This assumption is a subliminal driver of the current theories, but it is not based on any empirical evidence – it is something that may be true or false.

3) Also, if 'experience' and 'external world' come closer together, the need for some kind of internal interpreter is dramatically reduced. The transformation of the outside world into an internal representation or a virtual model is getting more and more obsolete.

4) The closer 'experience' and 'external world' get, the less their difference can be. This is not at all a trivial statement. On the contrary, this leads to the most difficult point to grasp: namely, that what people call consciousness 'is' the world people live in. It is not how the world appears, but what the world is.

In Chapter III the consequences of this new model will be discussed. First, however, the authors would like to give the reader the chance to understand HOW such an approach could work in practice with a construction sketch The key idea is that the body and the world are just two pieces of the same physical system and that what the authors call the mind is a physical process that requires both pieces to take place. Body and world

[2] Of course, these models do not rule out the importance of the external environment to allow the development of internal structures. Indeed, they consider it as necessary for a healthy brain development to continuously interact with the environment. However, once the required neural connections are in place, the mind is taken to be an internal phenomenon. Dreams and hallucinations are constantly quoted as obvious cases.

are interlocked gear wheels and the consciousness turns them. A schematic description of how the coupling between body and world works is sketched in Figures 1, 2 and 3.

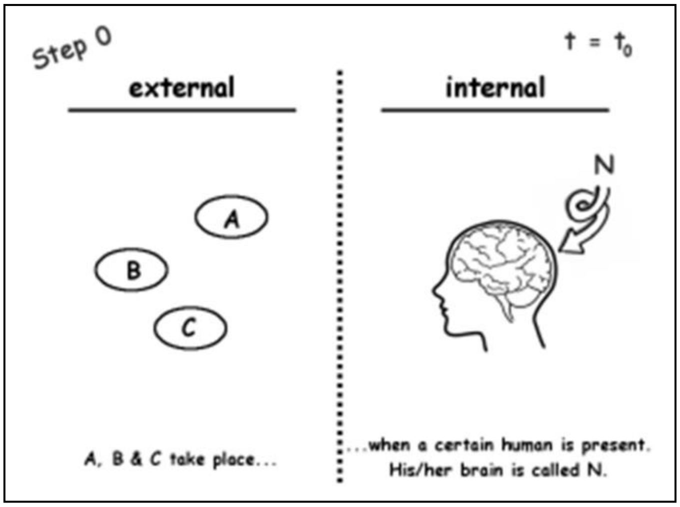
Step 0
t = t0
external
internal
N
A
B
C
A, B & C take place...
...when a certain human is present.
His/her brain is called N.

Fig. 1. Step 0 - before perception, there is no external object, as one perceives it. There are smaller and scattered physical phenomena, which are not the target of any normal experience.

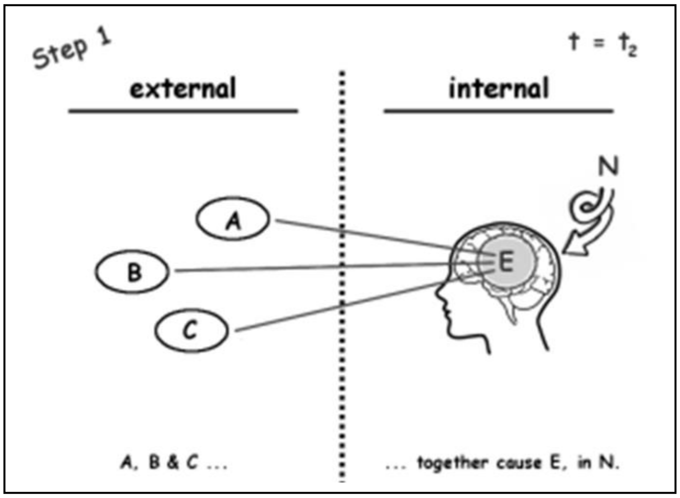

Fig. 2. Step 1 - because of the presence of a certain neural structure inside a body with the proper sensor apparatus, the scattered external phenomena produces a joint effect inside one's brain.

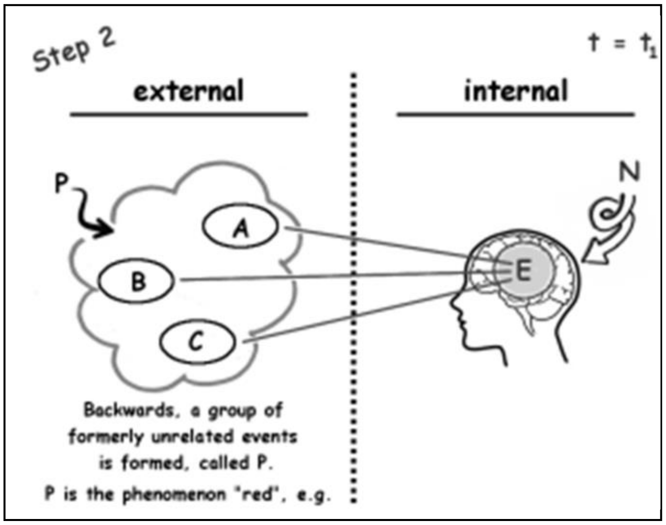

Fig. 3. Step 2 - events that are responsible for the occurrence of a joint effect are a joint cause – they become a whole. One's experience is the process P which takes place thanks to both: to one's brain and by A, B, and C.

Below the different steps are explained in detail:

- In the outside[3] world, there are physical scattered events. Let us call them A, B, C. They are located in time and space. A, B, and C do not have anything in common – each take place on their own.
  - For instance, A may be a certain light ray with a certain frequency emitted at a certain time in a certain location.
- In a nearby body, a healthy brain (call it N), through the causal connection provided by sensorial paths, is affected by these three events A, B, and C. This is not an obvious step. What now happens is that the neural structure – thanks to various neural learning mechanisms – allows the fusion of A, B, and C to create the causal circumstances[4] that would allow the fusion itself to act as a cause for further interaction. A whole is born where P is called the fusion of A, B, and C.
- E is the joint effect of P and takes place inside the brain N.

[3] Outside and inside are used to refer to physical events inside and outside one's body.

[4] In causal terms, one may distinguish between 'the cause and that without which the cause would not be a cause' ([20], p. 119). The former may be taken to be an event actually occurring while the latter may be just a state of affairs. The latter may be formalized in terms of conditions G such that $P \wedge G \rightarrow E$ (which may unfolded in three conditionals $P \wedge \neg G \rightarrow \neg E$, $\neg C \wedge G \rightarrow \neg E$, and $\neg C \wedge \neg G \rightarrow \neg E$.

  - As a simple analogy consider the 'distributed key' of an atomic rocket launch system: to launch it, two different 'keys' have to be turned at the same time (in movies, usually one of them always refuses to work!). However, considering the matter more carefully, there are not two keys, there is just one key in two pieces, and to be even more precise, the two pieces are not even keys since, when alone they do not unlock anything.

- Because the fusion P of A, B, and C causes E, these three originally independent events 'become' a whole in any practical and sense – things are defined in causal terms and not in ideal terms. P has not existed until E occurred. E is inside the body, while P remains outside of it.

  - Going back to the Rocket launcher analogy, before being put into the 'lock', both 'key pieces' are as unrelated as A, B, C. After being inserted, they, together with the lock, from a group and form a 'lock-and-key principle' which is a construction partially internal and partially external. The separation between them becomes purely conceptual – physically and causally they are a whole.

- Finally, the hardest bullet to swallow, namely the temporal order. No, the authors have NOT confused the indices: E takes place at a time t_2, while P occurred at a time t_1, with $t_1<t_2$. Why that – and does that mean that E is changing its past? In some sense yes. At least it is changing the causal role of the past.

The last point merits some further consideration. It is not the intention to invoke some sort of retro-causation that moves backward in time. Rather, the point shows something that should not come as a surprise: Physical phenomena are extended in time. This means that they get to completion within time. Therefore, when something begins to unfold, its nature is not wholly defined until it reaches some natural ending in causal terms. Nothing goes backward in time – the past is, of course, past.

However, what the past was may well be defined by what happens in the present. Using the above time indexes, there is no need to suppose that anything is going from t_2 to t_1. However, there is no harm either in considering that what the world at time t_1 was (that is, A, B, and C) changed after t_2 (that is, the fusion P takes place). If one considers a physical phenomenon as something extended in time, then it may well be the case that what happens along such an extension redefines the structure of the phenomenon.

To understand the temporal structure of the proposed causal sequence one has to take a closer look. At time $t_1<t_2$, the actual cause P has not yet happened. Until t_2, P has not yet produced any effect. Thus, from a physical standpoint, P has not yet existed. One may put the situation in these terms – until t_2 P's existence is not causally any different from P's absence. Then E takes place at t_2. Things have changed. P is now the actual cause of an event E that might have not happened. Yet E happened and thus P was its actual cause. Has this temporal sequence any effect to the temporal order of perceived events [21] [22]? Not necessarily, as a matter of fact, the subjective temporal order of events depends on how subsequent cognitive processes exploit them. Furthermore, here the crucial issue is the internal physical and causal structure of a perceptual act rather than how the temporal order of different perceptual acts is experienced. Just to dispel any possible misunderstanding, neither the time t_2 is not the subjective time, nor the time t_1 is the objective time. Both t_1 and t_2 are physical times and they refer to when certain causal processes take place. The interval t_1-t_2 has no mandatory impact as to which order P is in temporal relation with other perceived events.

As a further example, let's get back to the case that was mentioned at the onset – namely conscious perception[5] of colour. If one applies the approach just sketched, a colour is a collection of scattered and otherwise separate physical properties until they produce a joint effect in one's brain (E)[6]. When they do so, the scattered wavelengths can be considered as a set of external phenomena (A, B, C, …), namely the colour red. What is hard to grasp is how to step from some scattered wavelengths to the impression of a colour. The answer is that it happens in exactly the same way as one comes from a bundle of 'whatsoever-pieces' to a key – the components themselves do not constitute a whole (in respect to colour) but the sum of them does. The pieces merge into one key IF AND ONLY IF they have the opportunity to do so in causal and actual terms, meaning that there is a 'suitable lock' around (certain interactions with the eye-brain system). Then the whole may take place – without that lock, nothing may happen. By doing so, the scattered events make the colour red happen. In this account, the colour red is the causal fusion of the set of incoming wavelengths. It is neither an internal impression nor a mental ink. Red is an external whole whose occurrence is made possible by causal coupling with the neural event (the joint effect E). A colour-blind person would not have the 'suitable lock' – and therefore would not be conscious of the phenomenon that normally sighted subjects call colour. In physical terms, if there was only a colour-blind person in a certain environment, the combination of physical phenomena getting to an end in a normally-sighted subject would never be able to produce a causal joint effect.

To recap, a causal notion of fusion may thus be put forward – any group of events X_i is fused if and only if there are the causal conditions for a further event E to take place. The idea is that a fusion takes place only if it is the actual cause of some event. Thus, a fusion is not an abstract entity, but a physical occurrence with its own causal efficacy. For any group of events, there is a fusion, whenever they produce together an effect. The structure of neural networks embedded into one's nervous and sensor systems are ideal in this respect – they are the causal circumstances that allow complex events in the environment to be the actual causes of some bodily events, hereafter integrated into the agent's behaviour. Thus, the fusion

[5] As the distinguished neuroscientist Semir Zeki once said, there is no such a thing as unconscious perception. However, in this case the authors prefer to be redundant rather than misunderstood.

[6] These properties may be quite an inhomogeneous set of actual physical properties such as the reflected colour spectrum, percentage of certain components, contrastive ratios among different areas, and so forth. For the sake of the example, just a set of wavelengths is considered.

takes place because of some neural event, which is the effect of external groups of events. What is fused, though, is not inside the body, but it is the group of external events. The actual cause, too, remains outside of the body.

Coming back to the steps 1-4 noted in Chapter II the main changes with respect to the standard view are recapped:

- The gap between experience and world … is gone since in the upper model, the external events and the internal perceptions become one.
- The human-centred view … is gone, since in this model, experience is driven partly internal und partly external to a physical body and it is constituted by physical events. The experience is internal to the physical system that underpins it, and it is external to one's body. The body is, of course, nothing but as a subset of a larger physical superset of processes taking place in time.
- Here, the word external does not mean being "projected" but just being physical outside of one's neural structure. Such a notion very strongly suggests that the physical underpinnings of mental states are made of physical events taking place outside of one's body – a standpoint sometimes dubbed phenomenal externalism or externalism about phenomenal content. The goal is to single out a physical event identical with one's experience without having to resort to any mentalistic notion such as content, character, interpretation, projection, reference, and so forth.
- The internal interpreter … is gone or at least not necessary any longer. It was enough to relocate our insights as to what the physical underpinnings of one's mind are. The consciousness of 'seeing red' (instead of seeing several scattered wavelengths) is the result of the fitting between key-parts and lock. Red is not a meaning associated to some internal representation – red is a physical phenomenon in one's physical environment.
- Consciousness 'is' the world people live in: to make an example, to see red is to be united with an external collection of physical phenomena, since experience takes places as a temporally and causally extended phenomenon that includes internal and external components.

In comparison to the traditional view previously mentioned, a new approach is fleshed out (key differences in italics):

- External events & a neural event
  - *form a key-lock-system,*
  - which is, therefore, *partly internal - partly external*.
- The external events produce a certain neural activity,
  - an event inside the human body
- and, as a result,
  - *an external actual cause has occurred.*

B. The consequences overall

For a moment, before raising the inevitable objections, consider this view as a tentative scientific hypothesis as to the physical nature of consciousness – a scientific hypothesis insofar as it puts forward a falsifiable hypothesis as to what the mind is. If this hypothesis had any merit, a few conceptual advantages are immediately obvious:

- First, **the hard problem of consciousness**: The hard problem of consciousness, introduced by David Chalmers [6], addresses the problem of explaining how and why one has qualia and phenomenal experiences such as pain, colours, taste etc. (incl. 'Why does awareness of sensory information exist at all?' 'And why is there a subjective component to experience?'). The presented approach sweeps away the premises on which the hard problem of consciousness is based on (and thus the hard problem itself). In short, the hard problem is based on the dustbin model of the conscious mind [23] [24] in which a set of features is relocated that have been eschewed from the physical world. The idea is that – according to Chalmers – the hard problem is not an unavoidable chasm in the structure of nature, but a false issue created by assuming wrong premises. The approach presented here addresses such premises and indeed suggests a different picture. The mind and the world would no longer be two incommensurable and indeed autonomous domains, but the same one under two different perspectives.
- Second, the **mind-body-problem**: overt and covert dualism would finally be overturned. Dualism is not just the straw man often depicted as the traditional substance dualism contrasting matter and soul or body and mind. There are also forms of dualism that suggest a juxtaposition between cognition and the brain [25] [26], sometimes dubbed Cartesian Materialism [27] [28]. There is no longer the need to differentiate the way in which things look to subjects and the way in which things are. There is just a flow of physical phenomena causally interconnected.
- Third, **exclusiveness**: being conscious of something is a 'private' event – but in contrast to the traditional interpretation, the privacy is no longer created by an internal individual interpreter. It is no longer an exclusive and unbridgeable metaphysical privacy. Rather, it is the kind of privacy that prevents two individuals from eating the same piece of cake. It is a notion akin to that considered by the philosopher Mark Johnston who considers privateness as the impossible to receive the same anti-flu shot [29]. The exclusiveness follows from the fact that the pieces fuse into one key only if there is a 'suitable lock' – the suitable brain of a conscious agent - around. The causal interaction between internality and external world links the observed object and its observer. Of course, in presence of two similar groups of events, two similar brains let similar fusions occur.
- Fourth, **location of consciousness**: it is possible to physically locate the (conscious) mind into the physical

world. The location is not some inside neural activity though. However, it is possible to pinpoint a certain physical process and consider whether such a process is identical to one's own experience of, say, a red patch. It is thus possible to resurrect the theory of identity in terms of broader physical processes and not just in terms of neural processes. The fact that consciousness takes place partially outside the body is not in contrast with the impression one may have to be located inside the own body. Nothing in our experience points to where our experience takes place, only to what our experience is. If someone cuts a finger, he or she does not feel a pain inside the brain, but rather a pain in the finger. By the same token, it is not necessary that the process has to be located within our body.

- Fifth, *the misperception issue*[7] – namely, the fact that apparently one may experience things that are not physically present, as it happens in the case of hallucinations or dreams e.g. – has to be dealt with differently. They are no longer the result of a somewhat 'hyper-creative' internal interpreter, but of an unusual connection with real features in one's environment since this component has been removed from the picture. First, it is important to realize that our dreams are just 'boring' recombinations of the basic components of our past, albeit reshuffled in possibly original ways, they are chimeric but not innovative [30] [31] [32] [33]. Second, it is important to realize that ALL perceptions require a temporal lag between the object and the neural activity, due to the velocity of information transportation. Combining these two insights leads to a possible and fairly simple explanation approach, namely, that dreams and hallucinations may just happen to be cases of very long and reshuffled perception of one's world. So, tentatively, this approach suggests that the stuff dreams (and consciousness) are made of is the same stuff the world is made of.

- Sixth, **tabula rasa**. According to this view there is no mental content distinct from a physical event (that may be part of one's body, of course). Thus, one may experience a red apple or an itch in the elbow, but one cannot experience a pure mental content that one may experience. This is a very physical view that rules out any immaterial or purely mental content. By the same token, at the very beginning, organisms cannot have any experience since, by definition, they have not yet been in contact with any physical phenomenon. This does not prevent, of course, that either new born infants or foetuses may have consciousness as long as 1) they have a working neural system and 2) they perceive external events through parts of their bodies (or their mothers' bodies). However, the approach presented here rules out any innate or purely mental content of experience.

III. Consciousness Opening New Perspectives for Artificial Intelligence

Can the outlined approach help in shaping and devising an architecture capable of consciousness? The authors believe that it can, because it suggests a causal structure of consciousness and thus something that may help in singling out relevant architectures in an artificial agent. For instance, the approach suggests that being conscious is not a matter of either having the right internal code [34] [35], or having a central global dashboard [36] [37], or processing information in a certain way [38]. The advantage of this proposal is that it allows for rather precise indications as to why the causal coupling between the environment and the agent ought to be realized. Of course, by itself, the approach does not provide a complete picture of how to implement an intelligent agent. Many other aspects – often already addressed and partially implemented in AI and robotics – must flank what is here suggested. In sum, the suggested approach to consciousness does not aim to be alternative to other approaches in AI or in robotics, rather it aims to tune them in a way that should be productive for consciousness.

A. From machine intelligence to machine consciousness

It may be useful to make a comparison between current attempts to implement intelligence and consciousness. The understanding of what intelligence is – or what it is not – fills book. The notion derives from the Latin verb 'intellegere' ('understanding', more literally 'to have the choice between', 'to read between' [39] [40]). As a scientific term originated in psychology, the concept of intelligence addresses the cognitive capabilities of an entity, usually a human being. In these general terms, the notion partially overlaps with the psychological notion of consciousness. This is also partially due to the fact that both notions (consciousness and intelligence) are mongrel concepts that encompass several vague and not entirely coherent aspects.

By and large, an agent with intelligence is often considered as divided into three central parts following each other:

1) *recognition of external changes*
- having **sensory** components in order to receive stimuli from the external environment

2) *information processing*
- being capable of processing the sensory data together with internal knowledge in order to adapt behaviour, **cognition**

3) *reaction*
- having the capability to interact with external environment, realized by **actuators**

[7] Whenever it was necessary to point to the autonomy of the mental with respect to the physical domain, the issue of misperception has been the battering ram of both philosophers and scientists. Dream and hallucinations appear as formidable evidence in favour of an inner world. However, this approach promises to locate in the physical surrounding a physical cause for ANY experience. All cases of conscious experience ought to be revisited as cases of (admittedly unusual) perception. The approach presented here honestly stands or falls on whether it will succeed to show that – perhaps surprisingly – whenever there is consciousness there is a physical phenomenon, which is the content of one's experience. The authors cannot do justice here to the problem of misperception by and large. However, one can flesh out a template of the strategy – namely to address each purported case of misperception and to revise it in terms of actual perception. (One of the authors is actually working on such an account for most cases of misperception, from hallucination to illusions, from aftereffects to direct brain stimulations.)

Hereby, the Latin verb 'cognoscere' translates into 'conceptualize, recognize'. Cognition comprises the processes of information processing within an intelligent actor ('he sensory input is transformed, reduced, elaborated, stored, recovered, and used). Cognitive processes are divided into conscious and unconscious ones, e.g., by far not all learning processes are conscious. From this argumentation chain – *from intelligence to cognition to consciousness* – it follows that consciousness plays an important role in the understanding of intelligence.

Even if the majority of research done in the field of intelligence is directed towards human intelligence, the upper description states clearly that intelligence is NOT a primacy of humans. Obviously, many animals have a certain form of intelligence – proof is already given by observing your pet cat – even if it may differ from the human. Interestingly, the scientific status of 'consciousness' in animals continues to be hotly debated even if it is obvious that most animals have a phenomenal consciousness including a sense of pain, colour recognition, temperature etc. As mentioned above, the confusion is partly due to the variety of conceptions of consciousness. Researchers from different fields include very different aspects into the concept: a) phenomenal consciousness, b) the capability of thinking (thinking, remembering, planning, expecting), c) self-consciousness (awareness of oneself), d) consciousness of uniqueness (of oneself and of others) etc. Whereas phenomenal consciousness is probably part of most animals, it is still unclear if at least highly developed animals as mammals dispose of additional types of consciousness [5]. So the research space may be unfolded according to two broad criteria; one related to the kind of agents (animal, human or machine) and the other related to the kind of cognition involved (sensori-motor skill, symbolic capability aka traditional intelligence, linguistic capability, consciousness).

So which interim conclusions can be deduced? Machine consciousness lies in the promising middle ground between the extremes of biological chauvinism (i.e., only brains are conscious) and liberal functionalism (i.e., any behaviourally equivalent functional systems is conscious) [41]. One of the most central concepts behind 'intelligence' and perhaps the most difficult aspect to grasp is clearly not restricted to humans. From that it follows quite naturally that when building a technological system with a somewhat 'authentic intelligence', consciousness will have to play its part. Phenomenal consciousness – that is. It remains to be seen whether new concepts to realize this aspect will lead to insights into other components of consciousness.

B. *Weak versus strong machine consciousness*

The traditional and historically outdated distinction between weak AI and strong AI results from two different requirements: on the one hand, it follows from researchers focusing on different goals (more 'practical' vs. more 'principal'). On the other hand, a comprehensive philosophical debate on the nature of intelligence is driving the debate, including its exclusiveness or non-exclusiveness for humans (or other biological systems), ethical aspects, and the general possibility of reconstructing real intelligence, just to mention a few important aspects.

Weak AI addresses the position of artificial intelligence in philosophy that machines can demonstrate human-like intelligence, but do not necessarily have a mind, mental states or consciousness. Contrarily, strong AI supposes that some forms of artificial intelligence can reason and solve problems[8] as opposed to just making the humans feel that the machines are intelligent. In short: a weak AI-capable agent seems to be intelligent whereas a strong AI-capable agent is intelligent.

Obviously, the philosophical question behind this distinction is strongly related to the problem of consciousness. From that, it is not surprising that some authors suggested the possibility to distinguish between weak and strong artificial consciousness [3] [5]. In analogy to the weak vs. strong AI debate, weak artificial consciousness aims to deal with agents that behave as if they were conscious, at least in some respects, whereas strong artificial consciousness tries to address the design and construction of 'truly' conscious machines. Thus, the distinction between weak and strong artificial consciousness mirrors the dichotomy between true conscious agents and 'as if' conscious agents.

Although the distinction between weak and strong artificial consciousness sets a temporary working ground [5], it suggests a misleading view in so far as it suggests that a concept for a 'weak artificial consciousness' will help to gain a 'first understanding' on what consciousness might be and how it could be realized. Since it misses indispensables for the understanding of cognition – namely experience, i.e. phenomenal consciousness – the concept will not be adequate to overcome 'the riddle': Skipping the 'hard problem' is not a viable option in the business of making conscious machines [42].

Another argument may be raised against the temptation of 'weak artificial consciousness via the easy way': in nature, the development of consciousness goes along with increased intelligence. Most animals are exhibiting behavioural signs at least of phenomenological consciousness, human beings have a phenomenological consciousness and 'above'. 'Evolutionary optimization' is the most powerful optimization known so far (even if it takes its time). Thus, it seems to be highly unlikely that natural selection took such a long way to provide us with consciousness if there was a way to get all the advantages of a conscious being without actually producing it. Of course, this does not mean 'proof' – but the authors cannot help but to sincerely doubt it.

IV. CONCEPTS FOR BUILDING A CONSCIOUS MACHINE

Now, can a machine gain consciousness – that is, strong artificial consciousness as described in the previous section? Why – or why not? Is consciousness not a property of natural system that may thus be, at least in principle, realized by another physical system? And if so, how can that be done? Armed with the absence of a theoretical reason to reject the practical possibility, this paper addresses this issue.

[8] Sometimes, the term 'artificial general intelligence' ('AGI') is used to address strong AI, in particular by science fiction writers and within the community of futurists.

In the following, the authors are not outlining a strong theoretical formulation. Also, they are not capable – at this point – to give ‘a full proof’ (in a strict sense). Rather, it is the intention to show the inherent potential in the given interpretation of consciousness (Chapter II): As long as consciousness is interpreted as an ‘internalistic’ concept, there would be no change in modelling it: It remains to be something like ‘internal interpreter, e.g. transforming 10.000 x 700 nm into “red”’. Nobody knows why and how, except from that it happens. The internalistic interpretation may be true but this would not help oneself to come closer to any understanding of the concept behind it. However, if consciousness is interpreted in the sense as the authors proposed (halfway between ‘internalistic’ and ‘externalistic’), then it could be realized (at least as a toy model) as will be explained in the following. Thus, one can start to understand it and try to run tests on it, and so on. So by proposing this possible solution, the authors will sketch a ‘lab scenario’. Here, promising off-the-shelf technologies are considered that may fill the bill if deployed in the proper way.

A. Preparations for a tentative architecture for a conscious agent

Currently, many robotic setups and architectures are the result of careful programming since designers aim to solve specific sensorimotor, relational, or logic issues. A classic example is offered by robotic feats like Robocup[9] where teams of robots exploit algorithms devised by their designers to compete together in a soccer match. Although their behaviours may be very clever it is not the result of real adaptation on a high-cognition level. Of course, there are some robots capable of learning new skills and to adapt to novel situation, at least to a certain degree. However, explicit attempts of integrating consciousness into a robots’ intelligence are rare, and so far no model has been exceedingly convincing.

Compared to current robotic agents, biological agents like mammals and humans show a totally different kind of adaptability to novel stimuli. Mammals are capable of dealing with totally unexpected environmental challenges for which they could not possibly have any kind of inborn solution. Furthermore, it is a fair bet to assume that the complexity of their neural structure largely exceeds their genetic blueprint. Most mammals are capable not only of learning how to achieve goals but also of learning what goals have to be pursued [43] [44] – which is an important issue in respect to consciousness. As it has been observed [45] [46], the cortex shows an almost universal capability of autonomously adapting to novel kind of stimuli: ‘The fact that humans can learn and adapt to problems that did not exist when the initial model (the neocortex) was created is proof of the generic nature of the mechanisms used by the human brain.’ [47]. Thus, it makes sense to look for very general approaches capable, albeit with possible shortcomings, to model a unified and common approach to all aspects of cognition.

Empirical evidence shows that mammals exhibit a very high degree of neural plasticity and cognitive autonomy [48] [49] [50] to the extent that it is fair to suppose that any part of the cortex might develop almost any cognitive skill. If this supposition were true, it would mean that the neocortex, and possibly the thalamocortical system, exploit some kind of rather general architectural principle, mainly independent of the kind of incoming data.

There have been various attempts in the past to devise a general cognitive architecture [47] [45] [46]. This paper makes yet another attempt and takes advantage of a rather simple idea: true autonomy entails teleological openness. By being teleologically open the authors mean that the system is capable of developing new goals autonomously on the basis of environmental conditions and stimuli [44].

B. Objectives and motivations of the architecture

What are the ideal features that a cognitive architecture should have in order to adapt to a partially unknown body and environment? On the basis of the available literature and the empirical evidence a series of key features and their justification may be listed:

- The architecture must be based on a very limited number of kinds for basic building blocks – each kind exploiting the same common structure. Thus, the description length of the architecture must be kept to a minimum.

- This basic module might be freely replicated in order to cope with multiple sensor modalities and demanding incoming stimuli. This should ensure scalability.

- The basic module has to be able to develop its own goals and to use them both for its own development and for interacting with other modules. This should allow developing intentionality and a tight environment-architecture coupling.

- In principle, adding further modules (constrained only by the system resources) should lead to an increase in performances. Once again, this is important for scalability.

An architecture with the above features should be able to adapt to unknown situations and with a minimum of predesign. Rather than specifying all the algorithms and their mutual relationships, the above approach suggests a recipe to build a cognitive architecture given a body and an environment. Such a recipe is a lot less demanding in terms of description and a priori knowledge than a detailed plan. Furthermore, a recipe of such a universal scope offers many more advantages in terms of adaptability and flexibility.

Thus, the architecture the authors are willing to implement must satisfy the following requirements:

- Structure:
  - ○ it must be scalable
  - ○ it must be adaptable
  - ○ it must take advantage of memory more than speed
  - ○ it must be hierarchical
- Capabilities:

[9] http://www.robocup.org/

  - it must take into account the whole history of the system
  - it must develop fine grained new goals
  - it must develop overarching goals emerging out of the finer structure
- Additional Do’s and Don’ts:
  - it must not rely on explicit algorithms
  - it might have a limited number of more specialized versions of the same elementary block (for fine tuning, better performance, and optimization)
  - it must be coherent to what one knows about the biological structure of a mammalian brain

C. Combining multi agent systems with genetic algorithms

A tentative approach might be to realize a robot’s brain as a multi-agent system (MAS) once such an endeavour may find support by some additional key hypothesis as to the physical foundations of consciousness. MAS have been discussed already as a possible model to realize artificial brains, or as a model to explain the function of a brain (e.g. [51]). They have also been discussed as a possible extension of cognitive architectures as e.g. within the hybrid design oft CLARION (e.g. [52]). In computer sciences, MAS have become a very popular instrument during the last years when modelling complex heterogeneous distributed systems, which are organized ‘bottom up’. Their strength lies in predicting appearance of complex phenomena. The single agents have a certain degree of autonomy, they represent local views (in general, no agent has a full global view of the system, due to the complexity and the number of dynamically changing external dependencies), and they work decentralized (‘no master brain’). Topics where multi-agent systems are used include in particular the modelling of social and/or cooperative structures. Multi-agent system may be one of the key architectural principles necessary for a conscious mind.

Taking the new approach to consciousness as described in Chapter II as basis, in such a MAS each software agent would represent one ‘conscious-lock’ to a certain key, an external phenomenon. Thus, the resulting robotical brain would be conscious of the external events it has the appropriate locks for it, and the mechanism of building this consciousness would be exactly the same as for the human brain. So, the tentative idea is that MAS could offer the necessary architectural backbone for a conscious mind and that, once tuned to satisfy to some specific requirement, may be indeed the workable tool to begin designing a new kind of cognition.

At least three questions pose themselves immediately:

- **Complexity:** One may argue that by this approach, only a small number of locks can be realized due to the enormous programming effort needed otherwise.
- **Specification**: An even harder objection might be that in this way, the programmer may tend to mainly ‘imitate’ the human consciousness but does not develop one which is appropriate for the given robot with a certain form, function and so on.
- **Proof**: A third difficult point is the answer to the question as how one would like to *prove* that a certain robot really has a consciousness in a strong sense (compare Chapter IV.B).

To tackle all three problems with one approach, optimization algorithms have to be integrated, allowing to improve the multi agent system during runtime. Here, due to their ‘closeness’ to the underlying problem (a developing brain), genetic algorithms might form a natural choice: The ‘consciousness-locks’ have to be specialized to species, their mode of living, and the challenges presented to them[10]. Their special characteristics are probably not the result of some kind of ‘biological master plan’ for all living beings, but the result of a species-exclusive evolutionary process, which over millions of years has favoured individuals which are better adapted to their environment than others. In this understanding, the consciousness-lock (realized through multi agents) would be subject to the same evolutionary process, which has driven the whole design of a certain species, including the body shapes, motor skills, brain structure and the like. Genetic algorithms are precisely reproducing this kind of development.

The idea of using genetic algorithms to build a conscious brain is also one of the central design principles behind the cognitive architecture [53]. Genetic algorithms are a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. They are inspired by Darwin's theory about evolution. The idea was introduced in the 1960s by Ingo Rechenberg in his work ‘Evolution strategies’. His idea has been extended by many other researchers over the last decades. Today, they play an important role in many complex optimization problems and form an important concept for machine learning approaches. Genetic algorithms use mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. Over several generations, systems are optimized: Pairs of first generation solutions are taken and recombined. The ‘fittest’ solutions of this match are selected for the next generation. Mutations are used to enhance the genetic variety and thus, the overall solution space. The optimization goal – in nature given through environment and the corresponding challenges – is realized through a so-called fitness function which determines the quality of the solutions. Lately, combining multi agent systems with genetic algorithms has become popular in certain field as e.g. automated testing scenarios.

Assuming that consciousness is a capability of higher development of life forms, the following digest gives a first

[10] Consider the following example: Literature states that cats are somewhat colour-blind concerning the colour red, they see it as a shade of grey (whereas they have a perfect colour vision concerning e.g. green and blue). Well, the first finding is that one cannot be really sure about that, since one can only predict that from their eye anatomy – but what kind of ‘consciousness’ cats really have concerning the colour red is a totally different topic because at this point the design of key-lock-structure is unknown. It may be totally different to the human one. The second – and much more important – insight is that it might be less important for a cat to be capable of seeing red than for example for a bear: cats – being carnivores – do not have to differentiate between ripe and unripe apples since they would not eat them anyway. For a bear on the other hand – being omnivores consuming a large portion of fruits daily – the situation may show itself quite differently.

impression of the number of genetical iterations which are necessary to produce this kind of complex structures: About 3.5 billion years ago, the first life forms developed, monads with a very limited range of functionalities. Based on the development of genetical heredity through DNA molecules, advancements and progresses could be passed over to the next generation leading to first plants and simple animals which arose about 700 million years ago. 200 million years ago, mammals started to populate the earth. Humanoid life forms developed 70 million years ago and the homo sapiens species is only 500.000 years old. Even if it is difficult to tell from which stage in evolution consciousness has first entered the scene, referring to the current state of the art its development is part of a growing and more and more complex brain (ibidem). From this analysis, the reader might understand why the others consider genetic algorithms for the optimization job!

D. The practical side

How to proceed ‘practically’, meaning: how exactly are the genetic algorithms used to re-build the evolution of a robotical brain?

- Regarding the development of consciousness, one would start with a couple of given perceptions, each of them realized through a single agent, say regarding colours, temperature and the like which seem to play an important role for all living beings – a ‘basic set’ of conscious perceptions, so to say. This is ‘easy’ – and would form the ‘first generation solution’.
- Now, to make the system learn new conscious elements, the second preparative step is to place the robot in a certain challenging environment – meaning that certain tasks have to be given to him – in order to challenge his ‘consciousness enhancement’.
- Next, genetic algorithms come into play in order to produce variants of the robot’s ‘brain structure’: the single agents will be multiplied and altered through the means of the genetic algorithms. They will become multiplied, more complex and more varying. Some of the ‘new’ solutions will not survive as they do not particularly contribute to the tasks the robot is given. Others will survive as they enhance the robot’s capabilities to deal with its tasks. This, the resulting brain structure, will turn out having consciousness-locks which are complex and adapted to the individual needs of the specific kind of robot and its environment.

From that, there are two possibilities to infer that the robotic brain is really developed something like a consciousness by using genetic algorithms:

- First, the direct inspection would address the source code itself. Starting from a ‘basis set’ of agents in a MAS, the resulting system would consist of old and new software agents, the latter representing new conscious capabilities. The new code can be investigated, varied and different tests cases could be designed and analysed.
- A more ‘indirect’ inspection would be: A test scenario could be designed where consciousness for a certain perception area would definitely be necessary to solve a certain task. By design, this particular conscious perception would not be part of the basic conscious skills the system is starting with[11]. Now, if – after of a couple of some (more) ‘genetic rounds’ – the robotic brain would come up with new solutions for the given task, which definitely requires the enhancement of its consciousness, this would be a strong signal that it has developed a new perception in a certain area.

So, the combination of multi agent systems with genetic algorithms allows overcoming the upper mentioned problems:

- **Complexity**: Starting from a small number of locks, their expansion is realized by genetic algorithms which enhance the number of locks in order to optimize the system’s behaviour
- **Specification**: Since the optimization takes places in relation to a certain environment including specific challenges and particular tasks, the robotical brain develops a consciousness which is adapted to its own needs.
- ‘**Proof**’: The proof of whether a consciousness has been developed is not complete. However, on the more direct side, the investigation of the auto-generated source-code will deliver new insights. From the perspective of an indirect proof, it would address the development of a new conscious aspect rather than its existence. If a robot can adapt to a certain situation IF and only IF it develops a conscious perception for something that will be a strong hint that consciousness has been developed.

V. Summary & Outlook

Putting it all together: According to [54], there are three motivations to pursue artificial consciousness [55] [56] [57]:

1) implementing and designing machines resembling human beings (cognitive robotics);

2) understanding the nature of consciousness (cognitive science);

3) implementing and designing more efficient control systems.

Based on the presented new approach to consciousness lying between internalism and externalism, a possible technological design for a conscious machine has been

[11] Consider the following example: assume having a robot with colour consciousness as one of the basic components. This robot is part of a cooperative structure with other robots and humans, working together in a production line. Due to long geographical distances within factory, it would be absolutely necessary to be capable to ‘visualize temperature’, meaning to have a visual perception for extended areas within wavelengths between 700 nm and 1 mm (infrared). Here, humans would not be able to ‘see’ the wavelengths, since they have no consciousness for this wavelength area. However, the robot (using genetic algorithms on the multi agent system which is forming the ‘conscious part’ of the robotical brain) could develop a perception for this wavelength. By that, the robot might be capable to solve the task – finding a heat leak in an extended machinery – opposed to the human. Due to the fact that the brain is built in the upper described key-lock system that would mean that some kind of ‘new’ consciousness has been developed.

sketched addressing the upper mentioned goals. The approach is taking advantage of an architecture exploiting self-development of new goals, intrinsic motivation, and situated cognition. From a technological point of view, multi agent systems are used to model independent conscious perceptions. Genetic algorithms – as a subgroup of evolutionary algorithms – come into play to mimic the biological evolution of the brain's structure, thus allowing in general for adaptivity and scalability, and assuring some coherence to what humans know about the biological structure of brains of higher developed animals as e.g. mammals.

The architecture does not pretend to be either conclusive or experimentally satisfying. In the future, this rather sketchy outline of a cognitive architecture will be enhanced to a satisfying and more comprehensive architectural model. At this point, the authors will also integrate components of a cognitive architecture that has been partially implemented in previous setups [58] [59] [60]. The goal of the full architecture model is the implementation of the kind of development and environmental coupling through consciousness which was described in the previous sections.

On the other hand, up-to-date examples of highly distributed systems will be analysed in respect to their decision making processes (e.g., IBM's Watson which is operating on very distributed resources originally). These Systems show a new quality of artificial intelligence from which can be learned from: If high-developed intelligence includes consciousness, and if these big data oriented approached do produce results with a certain intelligence, than the interesting question arises whether these systems MUST have developed a certain consciousness, as part of their intelligence. If there is any merit in that, one could observe the emergence and the 'building' of consciousness in artificial system. Just by watching and interpreting, one could avoid arguing on the basis of biases and presumptions, bringing the whole debate back into the laboratories of natural sciences.

References

[1] C. Adami, "What Do Robots Dream Of?," Science, vol. 314, no. 5802, pp. 1093–1094, Nov. 2006.

[2] G. Buttazzo, "Can a machine ever become self-aware?," in Artificial Humans, R. Aurich, W. Jacobsen, and G. Jatho, Eds. Los Angeles: Goethe Institut, 2000, pp. 45–49.

[3] A. Chella and R. Manzotti, Eds., Artificial Consciousness. Imprint Academic, 2007.

[4] D. Gamez, "Progress in machine consciousness," Consciousness and Cognition, vol. 17, no. 3, pp. 887–910, Sep. 2008.

[5] O. Holland, Ed., Machine Consciousness. Imprint Academic, 2003.

[6] D. J. Chalmers, in The Conscious Mind: In Search of a Fundamental Theory, Oxford University Press, U.S.A., 1996, pp. xvii, 414.

[7] J. Kim, Mind in a Physical World. Cambridge, Mass: MIT Press, 1998.

[8] T. Nagel, "What is It Like to Be a Bat?," Philosophical Review, vol. 83, no. October, pp. 435–50, 1974.

[9] S. E. Palmer, Vision science: photons to phenomenology. Cambridge, Mass.: MIT Press, 1999.

[10] F. Crick, Astonishing Hypothesis: The Scientific Search for the Soul. New York: Touchstone, 1994.

[11] C. Koch and G. Tononi, "Can Machines Be Conscious?," IEEE Spectrum, vol. 45, no. 6, pp. 47–51, Oct. 2008.

[12] A. Revonsuo, Inner Presence: Consciousness as a Biological Phenomenon. Cambridge, Mass: MIT Press, 2006.

[13] Editors of Scientific American Magazine, The Scientific American Book of the Brain. Lyons Press, 1999.

[14] A. Damasio, Feeling of What Happens: Body and Emotion in the Making of Consciousness. New York: Harcourt Brace & Company, 1999.

[15] G. M. Edelman and G. Tononi, A Universe Of Consciousness: How Matter Becomes Imagination. London: Allen Lane, 2000.

[16] W. James, The Principles of Psychology. New York: Henry Holt and Company, 1890.

[17] B. Kuipers, "Drinking from the Firehose of Experience," Journal of Artificial Intelligence in Medicine, vol. 44, no. 2, pp. 155–170, 2008.

[18] R. Kurzweil, How to Create a Mind: The Secret of Human Thought Revealed. New York: Viking Adult, 2012.

[19] J. R. Searle, Minds, Brains, and Science. Harvard University Press, 1984.

[20] S. Yablo, "Advertisement for a Sketch of an Outline of a Prototheory of Causation," in Causation and Counterfactuals, J. Collins, N. Hall, and L. A. Paul, Eds. Cambridge, Mass: MIT Press, 2004, pp. 119–137.

[21] D. C. Dennett and M. Kinsbourne, "Time and the observer: The where and when of consciousness in the brain," Behavioral and Brain Sciences, vol. 15, no. 02, pp. 183–201, 1992.

[22] R. Flach and P. Haggard, "The cutaneous rabbit revisited," Journal of Experimental Psychology: Human Perception and Performance, vol. 32, no. 3, pp. 717–732, 2006.

[23] M. Cook, "Descartes and the Dustbin of the Mind," History of Philosophy Quarterly, vol. 13, no. 1, pp. 17–33, 1996.

[24] S. Shoemaker, "Qualities and Qualia: What's in the Mind?," Philosophy and Phenomenological Research, vol. 50, pp. 109–131, 1990.

[25] M. R. Bennett and P. M. S. Hacker, Philosophical Foundations of Neuroscience. Malden, Mass: Blackwell, 2003.

[26] R. Manzotti and Moderato, "Is Neuroscience Adequate As The Forthcoming 'Mindscience'?," Behavior and Philosophy, vol. 38, pp. 1–29, 2010.

[27] J. Dewey, Experience And Nature. Chicago: Open Court, 1925.

[28] W. T. Rockwell, Neither Brain nor Ghost. Cambridge, Mass: MIT Press, 2005.

[29] M. Johnston, "Appearance and Reality," in The Manifest, Princeton, NJ: Princeton University Press, 2002.

[30] C. S. Hurovitz, S. Dunn, G. W. Domhoff, and H. Fiss, "The Dreams of Blind Men and Women: A Replication and Extension of Previous Findings," Dreaming, vol. 9, no. 2–3, pp. 183–193, 1999.

[31] N. H. Kerr and G. W. Domhoff, "Do the Blind Literally 'See' in Their Dreams? A Critique of a Recent Claim That They Do," Dreaming, vol. 14, no. 4, pp. 230–233, 2004.

[32] A. Revonsuo and C. Salmivalli, "A content analysis of bizarre elements in dreams," Dreaming, vol. 5, no. 3, pp. 169–187, 1995.

[33] E. Schwitzgebel, C. Huang, and Y. Zhou, "Do we dream in color? Cultural variations and skepticism," Dreaming, vol. 16, no. 1, pp. 36–42, 2006.

[34] P. S. Churchland and T. J. Sejnowski, "Neural Representation and Neural Computation," Philosophical Perspectives, vol. 4, pp. 343–382, 1990.

[35] F. Tong and M. S. Pratte, "Decoding patterns of human brain activity," Annu Rev Psychol, vol. 63, pp. 483–509, 2012.

[36] B. J. Baars, "In the Theatre of Consciousness: Global Workspace Theory, a Rigorous Scientific Theory of Consciousness," Journal of Consciousness Studies, vol. 4, no. 4, pp. 292–309, 1997.

[37] M. Shanahan, Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds. Oxford University Press, USA, 2010.

[38] G. Tononi, "An information integration theory of consciousness," BMC Neuroscience, vol. 5, no. 42, pp. 1–22, 2004.

[39] D. J. Chalmers, "Facing Up to the Problem of Consciousness," Journal of Consciousness Studies, vol. 2, no. 3, pp. 200–219, 1995.

[40] R. Pfeifer and C. Scheier, Understanding Intelligence. Cambridge, MA, USA: MIT Press, 1999.

[41] R. Jackendoff, Consciousness and the Computational Mind. Cambridge, Mass: MIT Press, 1987.

[42] R. Manzotti, "Is Consciousness Just Conscious Behavior?," Int. J. Mach. Conscious., vol. 3, no. 2, pp. 353–360, Dezember 2011.

[43] R. Manzotti, F. Mutti, G. Gini, and S.-Y. Lee, "Cognitive Integration through Goal-Generation in a Robotic Setup," in Biologically Inspired Cognitive Architectures 2012, A. Chella, R. Pirrone, R. Sorbello, and K. R. Jóhannsdóttir, Eds. Springer, 2013, pp. 225–231.

[44] R. Manzotti and V. Tagliasco, "From behaviour-based robots to motivation-based robots," Robotics and Autonomous Systems, vol. 51, no. 2–3, pp. 175–190, Mai 2005.

[45] D. George and J. Hawkins, "Towards a Mathematical Theory of Cortical Micro-circuits," PLoS Comput Biol, vol. 5, no. 10, p. e1000532, Oktober 2009.

[46] J. Hawkins and S. Blakeslee, On Intelligence. New York: Times Books, 2004.

[47] D. George, "How the Brain Might Work: A Hierarchical and Temporal Model for Learning and Recognition," Stanford University, Stanford, CA, USA, 2008.

[48] J. Sharma, A. Angelucci, and M. Sur, "Induction of visual orientation modules in auditory cortex," Nature, vol. 404, no. 6780, pp. 841–847, Apr. 2000.

[49] J. Sharma, V. Dragoi, J. B. Tenenbaum, E. K. Miller, and M. Sur, "V1 Neurons Signal Acquisition of an Internal Representation of Stimulus Location," Science, vol. 300, no. 5626, pp. 1758–1763, Jun. 2003.

[50] M. Sur, P. E. Garraghty, and A. W. Roe, "Experimentally induced visual projections into auditory thalamus and cortex," Science, vol. 242, no. 4884, pp. 1437–1441, Dec. 1988.

[51] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, Principles of Neural Science, Fifth Edition, 5 edition. New York: McGraw-Hill Medical, 2012.

[52] R. Sun, "The CLARION Cognitive Architecture: Extending Cognitive Modeling to Social Simulation," in Cognition and Multi-Agent Interaction, R. Sun, Ed. Cambridge University Press, 2005, pp. 79–100.

[53] B. Goertzel and D. Duong, "OpenCog NS: A Deeply-Interactive Hybrid Neural-Symbolic Cognitive Architecture Designed for Global/Local Memory Synergy," in AAAI Fall Symposium: Biologically Inspired Cognitive Architectures'09, 2009, vol. FS-09–01.

[54] R. Sanz and C. Hernández, "Towards Architectural Foundations for Cognitive Self-aware Systems," in Biologically Inspired Cognitive Architectures 2012, A. Chella, R. Pirrone, R. Sorbello, and K. R. Jóhannsdóttir, Eds. Springer, 2013, pp. 53–53.

[55] J. Bongard, V. Zykov, and H. Lipson, "Resilient Machines Through Continuous Self-Modeling," Science, vol. 314, no. 5802, pp. 1118–1121, Nov. 2006.

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

[57] R. Sanz, I. López, M. Rodríguez, and C. Hernández, "Principles for consciousness in integrated cognitive control," Neural Networks, vol. 20, no. 9, pp. 938–946, Nov. 2007.

[58] R. Manzotti, L. Papi, and S.-Y. Lee, "Does radical externalism suggest how to implement machine consciousness?," in Biologically Inspired Cognitive Architectures 2011, A. Samsonovich and K. Jóhannsdóttir, Eds. Amsterdam: IOS Press, pp. 232–240.

[59] R. Manzotti, "A Process-Based Architecture for an Artificial Conscious Being," in Process Theories: Crossdisciplinary Studies in Dynamic Categories, J. Seibt, Ed. Springer Netherlands, 2003, pp. 285–312.

[60] R. Manzotti, "Machine Free Will: Is Free Will a Necessary Ingredient of Machine Consciousness?," in From Brains to Systems, C. Hernández, R. Sanz, J. Gómez-Ramirez, L. S. Smith, A. Hussain, A. Chella, and I. Aleksander, Eds. Springer New York, 2011, pp. 181–191.