Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 40 (2016) 172–185

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

A causal foundation for consciousness in biological and artificial agents

Action editor: V. G. Honavar

Riccardo Manzotti [a,b], Sabina Jeschke [c,*]

[a] Fulbright Visiting Scholar, Massachusetts Institute of Technology, Department of Linguistics and Philosophy, MA, United States
[b] IULM University, Milan, Italy
[c] Institute Cluster IMA/ZLW & IfU, Faculty of Mechanical Engineering, RWTH Aachen University, Aachen, Germany

Received 28 August 2014; accepted 8 November 2015
Available online 14 January 2016

Abstract

Traditional approaches model consciousness as the outcome either of internal computational processes or of cognitive structures. We advance an alternative hypothesis – consciousness is the hallmark of a fundamental way to organise causal interactions between an agent and its environment. Thus consciousness is not a special property or an addition to the cognitive processes, but rather the way in which the causal structure of the body of the agent is causally entangled with a world of physical causes. The advantage of this hypothesis is that it suggests how to exploit causal coupling to envisage tentative guidelines for designing conscious artificial agents. In this paper, we outline the key characteristics of these causal building blocks and then a set of standard technologies that may take advantage of such an approach. Consciousness is modelled as a kind of cognitive middle ground and experience is not an internal by-product of cognitive processes but the external world that is carved out by means of causal interaction. Thus, consciousness is not the penthouse on top of a 50 stores cognitive skyscraper, but the way in which the steel girders snap together from bottom to top.

Keywords: Machine consciousness; Consciousness; Cognitive architecture; Externalism; Situated cognition

1. Introduction

Is it possible to devise a general architectural principle that might lead an artificial agent to exploit what is called consciousness in human beings and various animals? In principle, there is no reason why such a feat should not be accomplished – at least if consciousness is a natural feature of certain biological beings as we believe it to be. However, it is still unclear what consciousness is and what the necessary and sufficient conditions for its occurrence are. Yet, consciousness appears to be a universal feature of all biological agents above a certain level of cognitive development. This is neither a sure proof that consciousness is a mandatory feature of any superior cognitive system nor that it is replicable by machine. However, it is a very strong evidence in favour of its usefulness and it physical nature. In this paper, we have three goals. First, we address a novel hypothesis about the physical conditions that lead to consciousness. Second, we show tentatively how such a condition can be exploited in artificial beings. Third, we discuss theoretical and technical consequences of the suggested approach.

In a scientific and technological context, whatever formulation of consciousness one adopts, *consciousness should*

* Corresponding author at: Institute Cluster IMA/ZLW & IfU, Faculty of Mechanical Engineering, RWTH Aachen University, Dennewartstraße 27, Aachen, Germany. Tel.: +49 2418091110.
E-mail addresses: riccardo.manzotti@gmail.com (R. Manzotti), sabina.jeschke@ima-zlw-ifu.rwth-aachen.de (S. Jeschke).

http://dx.doi.org/10.1016/j.cogsys.2015.11.001
1389-0417/

not come at the end of the day as an addition to an otherwise working description of a physical system. Since natural selection has singled out conscious agents capable of coping with the world, consciousness must make a not negligible physical difference. Thus, it must be identical to some physical structure or process. Of course, consciousness may be either the outcome of a process or may be embodied from the beginning as a consequence of some fundamental structural principle. While most approaches to machine consciousness privilege the former option, here we take into consideration the latter. The problem, thus, is *how to devise a model of the agent in which consciousness is not a nuisance but a natural and efficient way to organise its structure*. To do so, we believe, it is necessary to outline a novel theoretical model of the causal coupling between agent and environment. Consciousness is not some unexpected internal (and somehow magical) property but rather the expression of the physical structure of the agent and the way in which it couples with the external world.

In a nutshell, in this paper, we put forward a hypothesis as to what consciousness is in terms of simple causal structure, rather than suggesting that consciousness is the outcome of complex emergent neural/cognitive/computational processes. The key question is whether it is possible to single out structural principles that could be replicated in artificial beings so that they could show something akin to human and animal consciousness.

Coherently with such a goal, it is fair to maintain that a huge amount of empirical evidence suggests that consciousness does not seem to be a local capacity – such as language or face recognition – but rather the expression of a general structural principle shared by all neural areas (and not necessarily contained by them). While it is practically possible to switch off many specific mental capabilities without compromising consciousness, there is no evidence that it is possible to switch off consciousness without shutting down all other functions. When consciousness is severely impaired the subject is usually able to behave in an automatic and stereotyped mode such as during certain kind of epileptic seizures or during the anecdotal night drive (Armstrong, 1981). However, the subject is unable to exploit its capability to perform any new task.

While human beings perform amazingly in terms of sensory-motor capability and specific skills such as pattern recognition, discrimination, and planning, so far consciousness has remained elusive. There is no doubt that human beings – and likely many animals both mammals and not mammals (Edelman, Baars, & Seth, 2005; Seth, Baars, & Edelman, 2005) – exhibit and exploit a sort of integrated cognitive capability that is closely tied to the capability of having a rich experience of the world and one's perceptual states. However, it is still far from clear how such a capability arises in biological systems and to what extent it may be replicated and exploited by an artificial system.

Consciousness is a notorious mongrel concept (Block, 1995) insofar as it refers to various aspects of the mind such as phenomenal experience, unity of action, information integration, symbol grounding, autonomy, and development of intrinsic goals (Chalmers, 1996; Gamez, 2010; Harnad, 1990; Manzotti, 2010; Tononi, 2004). Yet, there are reasons – both of theoretical and empirical nature – that suggest that such a cluster of apparently loosely related mental feature is the outcome of a single architectural principle (Bullmore & Sporns, 2012; George, 2008; Manzotti & Chella, 2014; Mountcastle, 1997; Tononi, Edelman, & Sporns, 1998).

So far, there has been no machine able to show this kind of unified mental integration that a conscious being – be it a dog or a man – shows so uncompromisingly and smoothly. As an example of such astonishing capability to self-organise a unity of experiences one may consider the cases of severe brain damage, drug intoxication, sensori-motor impairment, neural forced rewiring, and genetic mutations (Hobson, 2002; Kahn & Krubitzer, 2002; Merker, 2007; Ramachandran & Hirstein, 1998; Sacks, 1970; Sur, Angelucci, & Sharma, 1999; Sur & Rubenstein, 2005). In all these cases, as long as the patient survives, the brain is able to self-adapt to conditions that could not have been part of its genetic blueprint. This is not to say that the brain is always able to cope with any damage to the extent that the subject is still operational. Often the damage overcomes the brain's resilience. Yet, it cannot be ignored that the brain, no matter what and as long as it has some residual capability, seems to be able to self-organize so that conscious experience occurs.

Here, we will focus on two aspects ubiquitous in conscious agents – namely the fact that conscious experience is always the experience of something and the fact that consciousness is always an integrated experience. Thus, content and unity. Is it possible to have a machine characterised by these two elements? And, finally, is their joint occurrence the hallmark of some structural architectural principle?

As regards these two aspects, the innovative rationale behind this paper is that, so far, in the still young literature about consciousness and even machine consciousness, two approaches have polarised most of the attention – namely computational approaches and action-oriented approaches. The former have focused their attention on special computations performed inside the agent (Baars, 1997; Koch & Tononi, 2008; Shanahan & Baars, 2005; Tononi, 2004; Tononi & Koch, 2008) – they believe that consciousness is some emergent property popping out of the computational processes inside the agent. The latter have focused on the sensori-motor coupling between an agent and its environment (Clark, 2013; Holland, 2004; Noë, 2004; O'Regan, 2012). Surely both approaches have their merit (Chrisley, 2003; Shanahan, 2010). Yet, it is possible that the nature of consciousness is the result of something that is physically more fundamental than the distinction between the inside and the outside of the agent. Consciousness might be the result of a very basic causal coupling between the body of the agent and the environment. It might be a basic principle *on the basis of which cognitive*

architectures develops. Consciousness might thus require to overcome the traditional distinction between *inner* computational processes and *outer* external stimuli. The goal of this paper is to address such a conceptual shift and to check whether it can be exploited to devise the basic requirements for a conscious architecture.

2. Material and methods

So far, many approaches to the issue of consciousness have had a weakness – if one would not know that certain physical activities are correlated with consciousness, the theory would have had no reason to predict its occurrence. (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Koch, 2004; Libet, 2005; Tononi, 2004; Zeki & Bartels, 1999).[1] Suppose to have a theory that postulates that, whenever a system performs a certain computation, consciousness occurs. What if consciousness did not occur? Would there be any difference in the outcome of the system? Of course not. In such theories, consciousness is added as a surprising and largely unexpected final effect, a bewildering *coup de théâtre*. Such approaches are unlikely to provide a satisfying explanation of what consciousness is.

Successful explanations in physics have a different structure. Consider a classic example. Thermodynamics is able to predict an increase in entropy and thus a variation of temperature given the initial conditions. Once you accept the premises of the theory, which can be checked experimentally, certain consequences follow. These consequences may thus be checked and their verification supports the theory.

In contrast, consider now a theory that suggests that consciousness is the result of a certain internal computation, for instance Tononi's theory of information integration (Tononi, 2004). This is an interesting hypothesis. But the point is that, if one starts assuming that consciousness is somewhat different from the physical activity taking place inside agents, consciousness cannot make any relevant difference afterwards. In other words, a computational process does not compute differently if it is labelled as "conscious". Likewise it does not make any difference whether such a computational process is correlated with a conscious process either running in parallel or emerging at the end of the computation. In technical terms, one may say that the simpler level (in this case the computational level) drains all the causal powers of additional levels (in this case consciousness) (Kim, 1998, 2003). So much for approaches that look for the neural/functional/computational correlates of consciousness.

[1] A necessary caveat: In this paper, whenever we criticize other approaches it is only to the extent that they are presented as an explanation of consciousness. Thus, for example, we have nothing either against Baars' theory of the global workspace as an excellent cognitive model, or Tononi's notion of integrated information as a mean to achieve unified representations. However, at the present state of research, we do not understand how these models could justify the emergence of consciousness *at the end of* a computational process or as the output of a cognitive module.

Likewise, consider an action-oriented form of explanation, for instance any stripe of enactivism (Noë, 2004). One considers an agent as a physical structure situated in a certain environment and the actions that such an agent performs. Once again, if one were able to provide a satisfactory theory of all the sensori-motor contingencies – that is, actions – between the agent and the world, why should one add anything like consciousness? The behaviour of the agent would be enough and the addition of consciousness would not and could not modify what is taking place.

In both cases, one could be reasonably tempted to get rid of the issue of consciousness altogether. However, this explanatory strategy (some form of eliminativism) would be unwise because (1) human beings seem to have something more than either sheer computation or sheer behaviour and (2) no purely computational agent or purely behaviour-based agent seems to have the degree of autonomy and adaptability that human and animal beings show. 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 feelings in other agents. In sum, consciousness appears to be a non-negotiable aspect of highly developed autonomous agents. The practical advantages that could result from its replication within an artificial being cannot be underestimated (Adami, 2006; Buttazzo, 2000; Gamez, 2008; Holland, 2003; Manzotti & Chella, 2009; Yourgrau, 2005). Thus, it is wise to reconsider the current state of research and consider alternative hypotheses about the basic requirements for consciousness.

Thus our goal is to outline a different starting point for the whole issue and to see whether such a different starting point could be exploited along the whole conceptual path that should have, as desirable outcome, either a prediction about the occurrence of consciousness in biological agents or the design of conscious artificial agents.

In short, what is the shift in perspective that we want to put forward? Basically, a theoretical framework that does not require the familiar distinction between the agent and the environment, between the internal cognitive processes and the external brute stimuli, between the internal states and the external causes, and so forth. Such distinctions, however useful, are suspicious because they presuppose a difference – inner vs. outer, inside vs. outside, computational vs. physical, input vs. output, process vs. action – that is the offshoot of assuming the existence of an agent. For instance, consider the notion of action, key for enactivists, which seems to set aside any vestigial form of dualism. Yet, is it possible to define actions irrespective of the existence of an agent? Is the shift between mere causal processes and action not akin to the gap that divides objects from agents? Is the notion of agent-less actions sound?

Thus, an alternative and radical hypothesis is put forward. Consciousness is not *something produced by an agent (no matter whether inside in terms of computational processes or outside in terms of actions), but rather it is the kind of basic causal structure between the body of the agent and the*

environment. Our goal is to outline such elementary but powerful causal structure and then, on the one hand, to show how it is exploited by existing biological beings and, on the other hand, to show how artificial agents can take advantage of it.

2.1. The approach

In this section, the objective is to provide an elementary causal building block of the interaction between the agent and the body that allows to single out experience as a kind of interaction. The resulting new concept of consciousness is directed to overcome the gap between 'experience' and 'physical world'.

One may ask why a causal building block and not either a computational, an information-related, or a functional building block? There are many possible reasons to avoid these descriptive levels of reality. We would like to name two of them. On the one hand, these levels run the risk to provide question-begging circular explanations of agenthood. It is dubious whether an agent-independent definition of computation is available – pace Chalmers and Tononi (Chalmers, 2011; Manzotti, 2012; Tononi, 2008). On the other hand, in a physical world the most fundamental kind of relation is a causal process – no matter how elusive it may be to logic modelling (Dowe, 2007; Paul & Hall, 2013; Reichenbach, 1958; Schaffer, 2014). Thus, if we want to build our architecture on top of fundamental notions, we better focus on something natural and physical such as causal processes.

The goal is to single out a physical and elementary building block that could be exploited to model the development of a (conscious) agent in a physical environment. Such a building block ought to be neutral with respect both to all mentalistic notions (such as computation, function, representation, and the like) and to the distinction between internal and external. Our bet is that by adopting such a neutral stance, the ensuing understanding of the agent structure will be greatly simplified.

First, however, we would like to give the reader the chance to understand *how* such an approach could work in practice. The key idea is that the body and the world are just two pieces of the same physical system and that what we call consciousness is an external cause that requires both pieces to take place. Body and world are interlocked.

Let us go through the different steps carefully as shown in Fig. 1:

- In the outside[2] world, there are physical scattered events (Fig. 1 top). 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 takes place on its own. For simplicity, assume they take place at time t_1.
  - For instance, A may be a certain light ray with a certain frequency emitted at a certain time and location and B and C surrounding colour spectra. Or A, B, C may be different words in a sentence, or three separate features of someone's face.

- In a nearby human body and after a finite amount of time (Fig. 1 middle), a healthy brain, through the causal connection provided by sensorial paths, is affected by the three events A, B, and C. Such an effect takes place at a time t_2 because necessarily the finite speed of causal processes requires a delay between external events and neural activity always occurs. This delay is due partially to the medium and partially to the complexity of the neural activity. The latter is usually a greater than the former. 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[3] that allow the fusion itself to act as a cause for further interaction. A whole is born. We call P the fusion of A, B, and C.
- N is the joint effect of P and takes place inside the brain (Fig. 1 bottom).
  - 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. However, considering the matter more carefully, there are not two keys, there is just one key in two pieces. The two pieces are not keys since, if alone, they do not unlock anything.

- Because the fusion P causes N, the three originally independent events (A, B, and C) 'become' a whole. Thus, things are defined in causal and not abstract terms. P does not exist until E occurs. N 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 system and form a 'lock-and-key principle' which is a construction partially internal and partially external. Their separation is only conceptual – physically and causally they are a whole.

- Finally, the hardest bullet to swallow, namely the temporal order. N takes place at a time t_2, while P occurs at a time t_1, with $t_1 < t_2$, meaning that the neural effect N in the brain happens after P comes into being even though P has not existed until N occurred. Does that mean that N changes its past? In some sense yes. At least

[2] We use outside and inside to refer to physical events inside and outside one's body.

[3] In causal terms, we may distinguish between 'the cause and that without which the cause would not be a cause' (Yablo, 2004, 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 $P \wedge G \rightarrow E$, $C \wedge G \rightarrow E$, and $C \wedge G \rightarrow E$.

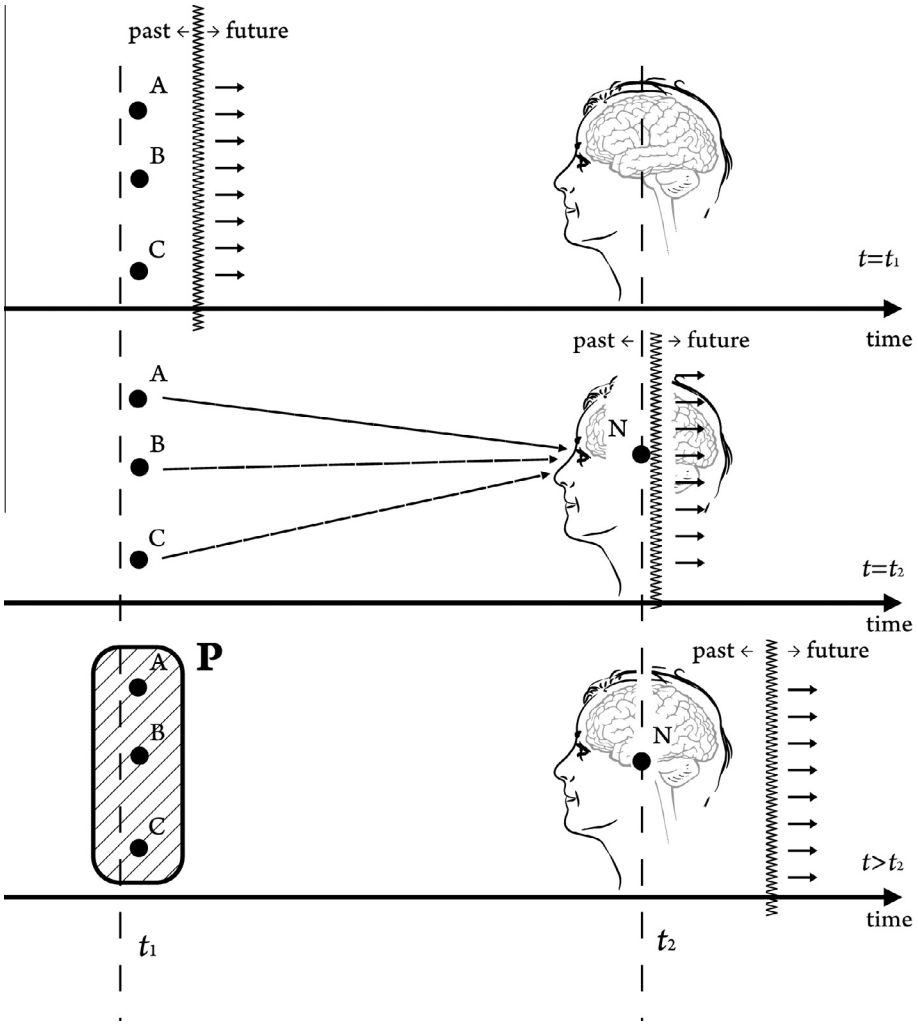

Fig. 1. A three-steps model of the proposed causal structure.

it changes the causal role of the past. But a better way to put it is the following. P remains unsaturated until N takes place. It is a bit like lottery tickets. The winning ticket is a winner only after it has been extracted. However, after it has been extracted, the lucky customer had bought the winning ticket from the beginning. So, macrophysical causes may be conceived as unsaturated functions that become saturated and thus complete only when their effects takes place.

The last point merits further considerations. There is no need to invoke any kind of retro-causation moving backward in time. Rather, the case shows that 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 now. Using the above time indexes, there is no need to suppose that anything goes backward from t_2 to t_1. However, there is no harm in considering that the world at time t_1 (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 time span redefines the structure of the phenomenon from the beginning.

As a further example, consider the conscious perception of colour. If we apply the approach just proposed, a colour is a collection of scattered and otherwise separate physical properties until they produce a joint effect in one's brain (N).[4] When they do so, the scattered wavelengths can be considered as a set of external phenomena (A, B, C, ...), say Tuscan red. What is hard to grasp is how to step from scattered wavelengths to the impression of a colour. The answer is that it happens in the same way as we come from a bundle of 'whatsoever-pieces' to a key – the components themselves do not constitute a whole (in respect to colour)

[4] These properties may be a inhomogeneous set of actual physical properties such as the reflected colour spectrum, the percentage of certain components, or the contrastive ratios among different areas. For the sake of the example, consider just a set of wavelengths.

until they produce a joint effect. 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 can happen. By doing so, the scattered events allow Tuscan red to take place. 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 possible thanks to causal coupling with the neural event (the joint effect N). A colour-blind person would not have the 'suitable lock' – and therefore would not be conscious of the phenomenon that standard sighted subjects call colour. If there were only a colour-blind person in a certain environment, the combination of physical phenomena getting to an end as a colour would never be able to produce a causal joint effect.

In a strong physical sense, where is the macroscopical object, which is one's experience (Fig. 2)? If a criterion of existence based on actual causation – namely to exist is to be the cause of some effect – is taken seriously, the object does exist only when it is the actual cause of an effect. By "actual" here we emphasise that the occurrence of the effect is mandatory. The effect happens thanks to the presence of the proper structure AND to many other contingent factors. For instance, one might be struck by a seizure and therefore be unable to complete the neural activity at the very last moment. Or, less dramatically, other processes may interfere and prevents neural processes from reaching completion as it is probably the case in many top down forms of blindness such as inattentional blindness and motion induced blindness (Bonneh, Cooperman, & Sagi, 2001; Simons & Chabris, 1999).

The new view is so different from the currently standard view that it is worth to make a quick comparison. The standard view in AI as well as in cognitive science is roughly the following: 1) a body capable of sensori-motor interaction with the environment; 2) external objects and their properties; 3) an input–output flow of information back and from such objects; 4) internal representations inside the body that – due to unknown reasons – are capable of producing the experience of external objects; 5) the brain as a kind of internal interpreter that gives meaning and "colour" to processed data; 6) consciousness as a property of the internal activity that is useful to control functional relations with the environment.

The current orthodoxy is based on the so-called 'internalistic models' – namely those models that take the mind to be a property of what takes place *inside* the neural system (Crick, 1994; Marr, 1982; Revonsuo, 2006; Tononi & Koch, 2008).[5] This view has its strength: putting some kind

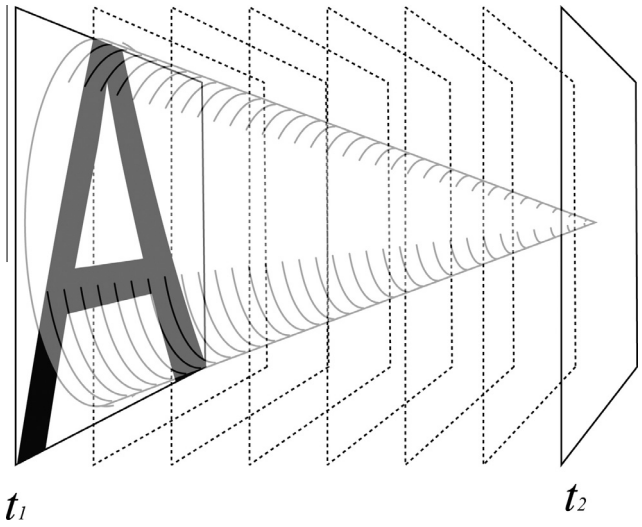

Fig. 2. Where is located in time an actual cause of a contingent process such as those relevant for everyday conscious perception? Is it located in a particular instant or rather it is spread in time between the occurrence of the components of the cause and the joint effect in one's brain?

of 'interpretation layer' between the individual and the outside world allows explaining why humans – all of them built alike – tend to be rather different in their behaviours, reactions and 'feelings'. However, there are also serious issues.

First, to believe that 'the internal interpreter', and 'neurons will do it "somehow"', is 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 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 such a model, consciousness neither fits the physical world nor its properties. To carry it to the extremes, that means that we constantly ignore the 'real world' by overwriting it with some internal 'fantasies', a sort of environment-driven virtual reality. 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 we experience – from pain to colour, from pictures to music – is nothing more than a product of our neurons, then a logical problem ensues: 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 not a cello? Or, to use an even catchier picture: 'If colours could not pop out of strawberries, how would they pop out of neurons?'

According to such a view, then consciousness is causally superfluous and the world is invisible. Furthermore, many

[5] Of course, these models do not rule out the importance of the external environment to guide the development of internal structures. Indeed, they consider necessary for a healthy brain to develop by means of continuous interactions with the environment. However, once the required neural connections are in place, the mind is taken to be an internal phenomenon.

explanatory terms seems to be circularly dependent on the notion of agent (and thus on that of consciousness). One may say that in the standard view, the homunculus in the brain (the ghost in the machine) has been dismissed because the whole brain has become the homunculus. Yet, the brain interprets, experiences, perceives, understands, wants, and so forth. Conveniently, the brain has become a kind of new homunculus that does everything that was once done either by the soul or by the immaterial Cartesian mind.

To overcome these obstacles and to achieve a foundation for an architecture that might be exploited by biological and artificial beings alike, our proposal dwells on a neutral ontology:

- The gap between experience and world ... is gone since external events and internal perceptions become a whole.
- The human-centred view ... is gone. 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.
- The internal interpreter ... is gone or at least no longer necessary. It was enough to relocate the physical underpinnings of consciousness. 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 we 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 which requires internal and external components. Red is external to one's body.

In short, the new approach may be recapped as follows.

- External events and a neural event form a key-lock-system that is neither internal nor external.
- The external events produce a certain neural activity whose goal is to allow a causal process between the body and the environment to take place.
- Because of the causal process between the neural activity and the environment, an actual external cause takes place.
- The external cause is both one's conscious experience and the external object/property.

In this way, because of the existence of an agent, the environment is partitioned so as to be made of a series of objects that are at the same time the experience and the object one is experiencing at any given moment.

2.2. Theoretical advantages

For a moment, before raising the inevitable objections, let us consider this view as a tentative scientific hypothesis about 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 has any merit, a few conceptual advantages will ensue:

- First, the hard problem of consciousness (Chalmers, 1996), addresses the problem of explaining how and why we have 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?'). With the presented approach, the core of the hard problem of consciousness is swept away. The mind and the world are no longer two incommensurable and indeed autonomous domains; they are the same one seen from two different perspectives.
- Second, the mind–body-problem. Overt and covert dualism is finally upturned. Dualism is not only the straw man of the traditional substance dualism contrasting either matter and soul or body and mind. There is also a form of dualism that suggests a juxtaposition between cognition and the brain (Bennett & Hacker, 2003; Manzotti & Moderato, 2010), sometimes dubbed Cartesian Materialism (Dewey, 1925; Rockwell, 2005). 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. The exclusiveness follows from the fact that the pieces fuse into one key only if there is a 'suitable lock' – a suitable brain and a working body – around. The causal interaction between internal 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 in the physical world. The location is not inside the neural system though. However, it is possible to pinpoint a certain physical cause and consider whether such a cause 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 might take place outside the body is not in contrast with the impression one may have to be located inside one's own body. Nothing in our experience points to where our

experience takes place, only to what our experience is. If we cut a finger, we do not feel a pain inside the brain; rather we feel a pain in the finger. By the same token, it is not necessary that the physical phenomenon that is our conscious experience is located inside our body.

- Fifth, the misperception issue[6] – namely, the fact that apparently we 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 unusual connections with real features in one's environment. It is important to realise 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 (Hurovitz, Dunn, Domhoff, & Fiss, 1999; Kerr & Domhoff, 2004; Revonsuo & Salmivalli, 1995; Schwitzgebel, Huang, & Zhou, 2006). It is also important to realise that all perceptions require a temporal lag between the object and the ensuing neural activity, due to the velocity of information transportation. Combining these two insights leads to a possible and fairly simple explanation, namely, that dreams and hallucinations might 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.

A way to visualise one of the main differences between this approach and other ones is given by the direction of the explanation (Fig. 3). In many approaches, first a theory of cognition (or something akin to it) is fleshed out and then a hypothesis is put forward about how a set of unconscious computation/processes/cognitive modules will produce consciousness. In contrast here a different stance is adopted. The key hypothesis is that consciousness is not the outcome (or an emergent property) of some internal activity. Consciousness is a way to describe the causal connection that holds between the external world and the agent's body. Thus, consciousness is not something that is concocted inside, but rather the way in which the interaction between the body (with all its neural structures) and the environment carves out a set of entities.

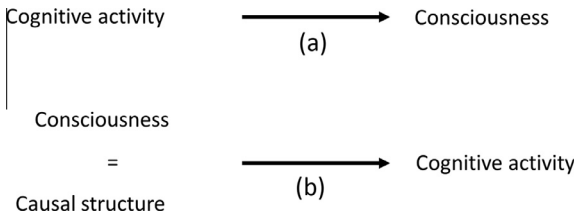

Fig. 3. The direction of the explanation in most standard cognitive approaches to consciousness (a) and in this case (b). Consciousness is at the root of the cognitive architecture.

3. Results and discussions

Now, can a machine gain consciousness – i.e., real consciousness?[7] To address this question we take advantage of the suggested fundamental causal structure and consider a series of promising cases and technologies.

In the following, we are not outlining a strong theoretical formulation. Also, we are not capable at this point to give 'a full proof' (in a strict sense). Rather, we want to show the inherent potential of the suggested causal model. As long as consciousness is interpreted as an 'internalist' concept, there would be no change in modelling it as an 'internal interpreter, e.g. transforming about 10.000 light waves of 700 nm into "red"'. Nobody knows why and how, except from that it happens. The internalist interpretation *may* be true, but this would not help us to come closer to any understanding of the concept behind. However, if consciousness is interpreted in the sense we have been proposing (halfway between 'internalist' and 'externalist'), then it could be realised (at least as a toy model) as we are going to explain. Thus, we can start to understand it, we can try to run tests on it, and so on. So by proposing this possible solution, we will sketch a 'lab scenario'. Here, we consider promising off-the-shelf technologies that may fit the bill if deployed in the proper way.

It is worth to stress that such a neutral causal account of consciousness is coherent with a notion of consciousness as cognitive middle ground – namely consciousness is not a special ingredients or the outcome of a special computational process but a basic causal structure underpinning the agents. Thus, to use a visual metaphor, *consciousness is not the penthouse on top of a 50 stores cognitive skyscraper, but it is the way in which the steel girders snap together*

[6] 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 a physical cause for any experience in the physical surrounding. 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. We cannot do justice here to the problem of misperception. However, we flesh out a template of the strategy – namely to address each purported case of misperception and to revise them in terms of 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.)

[7] The distinction between strong and weak machine consciousness mirrors that between strong and weak AI (Holland, 2003). Weak machine consciousness considers whether it is possible to build machines that behave as if they were consciousness. Strong machine consciousness ventures to consider the possibility of real conscious machines. We believe that skipping the 'hard problem' is not a viable option in the business of making conscious machines (Manzotti, 2011a).

from bottom to top. In fact, 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) (Jackendoff, 1987). 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 the goal of building a technological system with a somewhat 'authentic intelligence' requires consciousness to be part of the game, i.e., phenomenal consciousness. It remains to be seen whether new concepts will lead to insights into other components of consciousness such as self-consciousness or forms of higher order consciousness.

3.1. Tentative guidelines for a conscious artificial 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 can be listed:

- The architecture must be based on a very limited number of kinds of basic building blocks – each kind exploiting the same common structure. Thus, the description length of the architecture must be kept to a minimum.
- The 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 to develop 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 we are willing to implement must satisfy the requirement of being both scalable and adaptable. Furthermore, the architecture has to take into account the whole history of the system and it must be coherent to the current understanding of the biological structure of a mammalian brain. It might have a limited number of more specialised versions of the same elementary block (for fine tuning, better performance, and optimisation), but it must not rely on explicit algorithms.

These requirements are definitely compatible with the neuroscientific evidence collected from human beings and non-humans mammals. Surprisingly, these requirements are not met by most artificial architectures and AI agents. In fact, 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 the Robocup[8] 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 at 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 (Manzotti, Mutti, Gini, & Lee, 2013; Manzotti & Tagliasco, 2005) – which is an important issue in respect to consciousness. As it has been observed (George & Hawkins, 2009; Hawkins & Blakeslee, 2004), 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.' (George, 2008). 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 (Sharma, Angelucci, & Sur, 2000; Sharma, Dragoi, Tenenbaum, Miller, & Sur, 2003; Sur, Garraghty, & Roe, 1988) 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 (George, 2008; George &

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

Hawkins, 2009; Hawkins & Blakeslee, 2004). In this paper, we make yet another attempt. This time we want to take advantage of a rather simple idea: true autonomy entails *teleological openness*. By teleologically openness we mean that the system is capable of developing new goals autonomously on the basis of environmental conditions and stimuli (Manzotti, 2010; Manzotti & Tagliasco, 2005).

3.2. Not reinventing the wheel: combining multi agent systems and genetic algorithms

A tentative approach might be to realise a robot's brain as a multi-agent system (MAS) once such an endeavour may find support by some additional key hypothesis about the physical foundations of consciousness. MAS have been discussed already as a possible model to realise artificial brains, or as a model to explain the function of a brain (e.g. (Kandel & Squire, 2000)). They have also been discussed as a possible extension of cognitive architectures – e.g. within the hybrid design of CLARION. In computer sciences, MAS have become a very popular instrument during the last years when modelling complex heterogeneous distributed systems, which are organised 'bottom up'.

Taking the suggested approach to consciousness as guideline, in such a MAS each software agent would represent one 'conscious-lock' to a certain key, an external phenomenon. Thus, the resulting robotic 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; it may be indeed the workable tool to design a new kind of cognition.

At least three questions arise immediately:

- *Complexity*: One may argue that by this approach, only a small number of locks can be realised due to the enormous programming effort needed otherwise.
- *Specification*: An even stronger 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 about how we would like to prove that a certain robot really has a consciousness in a strong sense.

To tackle all three problems with one approach, optimisation algorithms have to be integrated, allowing to improve the MAS during runtime. Here, due to their 'closeness' to the underlying problem (a developing brain), genetic algorithms might form a natural choice: The 'con sciousness-locks' have to be specialised to species, their mode of living, and the challenges presented to them.[9] 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 (realised 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. Likely, genetic algorithms reproduces 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. Genetic algorithms are a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. Today, they play an important role in many complex optimisation 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 optimised: 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 optimisation goal – in nature given through environment and the corresponding challenges – is realised 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 life forms, the following digest gives a first impression of the number of genetic iterations which are necessary to produce this kind of complex structures: About approx. 3.5 billion years ago, the first life forms developed, monads with a very limited range of functionalities. The first plants and simple animals came up about 700 million and mammals about 200 million years ago. Humanoid life forms started to develop 70 million years ago, and we, the homo sapiens species, are only 150.000 years old. Even if it is difficult to tell from which stage in evolution consciousness

[9] Consider the following example: In the literature, we find 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 we cannot be really sure about that, since we 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 we don't know anything about the design of the key-lock-structure at this point. It may be totally different from ours. 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.

has first entered the scene, the reader might understand why we consider genetic algorithms to do the optimisation job!

How do we proceed 'practically', meaning: how exactly are we going to use genetic algorithms to re-build the evolution of a robotic brain?

Regarding the development of consciousness, one would start with a couple of given perceptions, each of them realised through a single agent, say regarding colours, temperature and the like – a 'basic set' of conscious perceptions. 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 altered, multiplied and become more complex. Some of the 'new' solutions will not survive as they do not particularly contribute to the robot's task. 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 has really developed consciousness by using genetic algorithms. First, the direct inspection would address the source code itself. Starting with 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 test cases could be designed and analysed. Second, a more 'indirect' inspection would be to use a test scenario that is designed in such a way that consciousness for a certain perception area would definitely be necessary to solve a certain task. This particular conscious perception would not be part of the basic conscious skills the system it started with. Now, if – after of a couple of '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 aforementioned problems:

- *Complexity*: Starting from a small number of locks, their expansion is realised by genetic algorithms which enhance the number of locks in order to optimise the system's behaviour.
- *Specification*: Since the optimisation takes places in relation to a certain environment including specific challenges and particular tasks, the robotic brain develops the right causal structure, 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 and if and only if it is in the right causal structure then that will be a strong hint that consciousness has been developed.

4. Conclusions

The just outlined causal structure, no matter how simple, suggests a kind of causal structure that may be exploited both by artificial and biological agents. As a matter of fact, there is a huge empirical evidence showing that biological beings take advantage of this kind of causal connection with their environment during their development (Aboitiz, Morales, & Montiel, 2003; Neisser, 1987; Thelen, 2000). Such a causal structure offers a physical basis to address the issue of consciousness since it suggests that experience might be the way in which the external object takes place thanks to the interaction with the body of the agent. In this regards this approach offers the following advantages:

(1) Experience is no longer something that the agent must concoct inside.
(2) Experience of an object and the object are, to a great extent, the same thing.
(3) Experience (and thus the mind) is no longer the outcome of an internal process but the very process taking place between the external object and the body of the agent.
(4) Experience is no longer the outcome of some problematic level of reality (information/computation/function).
(5) Experience might be modelled into an artificial agent by controlling the way in which the artificial agent interacts with the world.
(6) The abovementioned way to interact with the environment might be the right one to achieve the necessary autonomy and adaptability exhibited by conscious biological agents.

These elements are, in our opinion, the preconditions for devising an architecture capable of consciousness. They suggest the 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 having the right internal code (Churchland & Sejnowski, 1990; Tong & Pratte, 2012), or using a central global dashboard (Baars, 1997; Shanahan, 2010), or processing information in a certain way (Tononi, 2004). The advantage is that this proposal provides precise indications about how the causal

coupling between the environment and the agent ought to be realised. 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 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.

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. 'Evolutionary optimisation' is the most powerful optimisation 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 we cannot help but sincerely doubt it.

So far, practical attempts to devise and implement conscious artificial beings have been hampered both by technical difficulties and by theoretical uncertainties. While the first issue will likely be overcome in time and is shared by other equally formidable feats in AI, developing consciousness in an artificial agent has often appeared to be an impossible endeavour (a preposterous idea). It is true that a handful of researchers have nonetheless considered the possibility, yet there is still no consensus as to what should be achieved.

Based on the presented new approach to consciousness lying between internalism and externalism, a possible technological design for a conscious machine has been sketched addressing the abovementioned goals. The approach is taking advantage of an architecture exploiting self-development of new goals, intrinsic motivations, 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 we 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. Then, we will also integrate components of a cognitive architecture which has been partially implemented in previous setups (Manzotti, 2003; Manzotti, 2011b; Manzotti, Papi, & Lee, n.d.; Manzotti & Tagliasco, 2005). The goal of the full architecture model is the implementation of the kind of development and environmental coupling through consciousness we have described in the previous sections. However, the presented approach suggests a way in which a cognitive architecture may exploit the same kind of causal entanglement that in biological beings might be identical with consciousness – *an agent is conscious of a given environment to the extent that its cognitive structures are the result of the proper causal coupling with that environment* – namely that one's consciousness of an object might be nothing but the external object tightly coupled to the agent's body. The experience of something would be literally identical to that something. Consciousness would be situated in the environment in a very strong sense and it would also be a very physical phenomenon.

References

Aboitiz, F., Morales, D., & Montiel, J. (2003). The evolutionary origin of the mammalian isocortex: Towards an integrated developmental and functional approach. *The Behavioral and Brain Sciences, 26*(5), 535–552 (discussion 552–585).

Adami, C. (2006). What do robots dream of? *Science, 314*(5802), 1093–1094. http://dx.doi.org/10.1126/science.1135929.

Armstrong, D. M. (1981). What is consciousness? In J. Heil (Ed.), *The nature of mind* (pp. 55–77). Cornell University Press.

Baars, B. J. (1997). In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies, 4*(4), 292–309.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Malden, Mass: Blackwell.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247. http://dx.doi.org/10.1017/S0140525X00038188.

Bonneh, Y. S., Cooperman, A., & Sagi, D. (2001). Motion-induced blindness in normal observers. *Nature, 411*(6839), 798–801. http://dx.doi.org/10.1038/35081073.

Bullmore, E., & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience, 13*(5), 336–349. http://dx.doi.org/10.1038/nrn3214.

Buttazzo, G. (2000). Can a machine ever become self-aware? In R. Aurich, W. Jacobsen, & G. Jatho (Eds.), *Artificial humans* (pp. 45–49). Los Angeles: Goethe Institut.

Chalmers, D. J. (1996). In *The conscious mind: In search of a fundamental theory*. USA: Oxford University Press, pp. xvii, 414.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science (Seoul), 12*(4), 323–357.

Chrisley, R. (2003). Embodied artificial intelligence. *Artificial Intelligence, 149*(1), 131–150. http://dx.doi.org/10.1016/S0004-3702(03)00055-9.

Churchland, P. S., & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives, 4*, 343–382.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences, 36*(3) , 181–204. http://dx.doi.org/10.1017/S0140525X12000477.

Crick, F. (1994). *Astonishing hypothesis: The scientific search for the soul*. New York: Touchstone.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204–211. http://dx.doi.org/10.1016/j.tics.2006.03.007.

Dewey, J. (1925). *Experience and nature*. Chicago: Open Court.

Dowe, P. (2007). Causal processes. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/archives/fall2008/entries/causation-process/>.

Edelman, D. B., Baars, B. J., & Seth, A. K. (2005). Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition, 14*(1), 169–187. http://dx.doi.org/10.1016/j.concog.2004.09.001.

Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition, 17*(3), 887–910. http://dx.doi.org/10.1016/j.concog.2007.04.005.

Gamez, D. (2010). Information integration based predictions about the conscious states of a spiking neural network. *Consciousness and Cognition, 19*(1), 294–310. http://dx.doi.org/10.1016/j.concog.2009.11.001.

George, D. (2008). *How the brain might work: A hierarchical and temporal model for learning and recognition*. Stanford, CA, USA: Stanford University.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology, 5*(10), e1000532. http://dx.doi.org/10.1371/journal.pcbi.1000532.

Harnad, S. (1990). The symbol grounding problem. *Physica, D*(42), 335–346.

Hawkins, J., & Blakeslee, S. (2004). *On intelligence*. New York: Times Books.

Hobson, J. A. (2002). *The dream drugstore: Chemically altered states of consciousness*. MIT Press.

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

Holland, O. (2004). The future of embodied artificial intelligence: Machine consciousness? In F. Iida, R. Pfeifer, L. Steels, & Y. Kuniyoshi (Eds.), *Embodied artificial intelligence* (pp. 37–53). Berlin Heidelberg: Springer, Retrieved from <http://link.springer.com/chapter/10.1007/978-3-540-27833-7_3>.

Hurovitz, C. S., Dunn, S., Domhoff, G. W., & Fiss, H. (1999). The dreams of blind men and women: A replication and extension of previous findings. *Dreaming, 9*(2–3), 183–193. http://dx.doi.org/10.1023/A:1021397817164.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, Mass: MIT Press.

Kahn, D. M., & Krubitzer, L. (2002). Massive cross-modal cortical plasticity and the emergence of a new cortical area in developmentally blind mammals. *Proceedings of the National Academy of Sciences, 99*(17), 11429–11434. http://dx.doi.org/10.1073/pnas.162342799.

Kandel, E. R., & Squire, L. R. (2000). Neuroscience: Breaking down scientific barriers to the study of brain and mind. *Science, 290*(5494), 1113–1120. http://dx.doi.org/10.1126/science.290.5494.1113.

Kerr, N. H., & Domhoff, G. W. (2004). Do the blind literally "See" in their dreams? A critique of a recent claim that they do. *Dreaming, 14*(4), 230–233.

Kim, J. (1998). *Mind in a physical world*. Cambridge, Mass: MIT Press.

Kim, J. (2003). Blocking causal drainage and other maintenance chores with mental causation. *Philosophy and Phenomenological Research, 67*(1), 151–176.

Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Roberts & Company Publishers.

Koch, C., & Tononi, G. (2008). Can machines be conscious? *IEEE Spectrum, 45*(6), 47–51.

Libet, B. (2005). *Mind time: The temporal factor in consciousness*. Cambridge, Mass: Harvard University Press.

Manzotti, R. (2011a). Is consciousness just conscious behavior? *International Journal of Machine Consciousness, 3*(2), 353–360. http://dx.doi.org/10.1142/S1793843011000765.

Manzotti, R. (2012). The computational stance is unfit for consciousness. *International Journal of Machine Consciousness, 4*(2), 401–420. http://dx.doi.org/10.1142/S1793843012400239.

Manzotti, R., & Chella, A. (2009). Artificial consciousness. *Thorverton*.

Manzotti, R., & Chella, A. (2014). Physical integration: A causal account for consciousness. *Journal of Integrative Neuroscience, 13*(2), 403–427. http://dx.doi.org/10.1142/S0219635214400044.

Manzotti, R., Papi, L., & Lee, S.-Y. (n.d.). Does radical externalism suggest how to implement machine consciousness? In A. Samsonovich, & K. Jóhannsdóttir (Eds.), *Biologically Inspired Cognitive Architectures 2011*. Amsterdam: IOS Press.

Manzotti, R. (2011b). Machine free will: Is free will a necessary ingredient of machine consciousness? In C. Hernández, R. Sanz, J. Gómez-Ramirez, L. S. Smith, A. Hussain, A. Chella, & I. Aleksander (Eds.), *From brains to systems* (pp. 181–191). New York: Springer, Retrieved from <http://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15>.

Manzotti, R., & Moderato, P. (2010). Is neuroscience adequate as the forthcoming "Mindscience"? *Behavior and Philosophy, 38*, 1–29.

Manzotti, R., Mutti, F., Gini, G., & Lee, S.-Y. (2013). Cognitive integration through goal-generation in a robotic setup. In A. Chella, R. Pirrone, R. Sorbello, & K. R. Jóhannsdóttir (Eds.), *Biologically inspired cognitive architectures 2012* (pp. 225–231). Springer, Retrieved from <http://link.springer.com/chapter/10.1007/978-3-642-34274-5_40>.

Manzotti, R. (2010). A process-oriented framework for goals and motivations in biological and artificial agents. In R. Poli (Ed.), *Causality and motivation* (pp. 105–134). Frankfurt: Ontos-Verlag.

Manzotti, R. (2003). A process-based architecture for an artificial conscious being. In J. Seibt (Ed.), *Process Theories: Crossdisciplinary Studies in Dynamic Categories* (pp. 285–312). Netherlands: Springer, Retrieved from <http://link.springer.com/chapter/10.1007/978-94-007-1044-3_12>.

Manzotti, R., & Tagliasco, V. (2005). From behaviour-based robots to motivation-based robots. *Robotics and Autonomous Systems, 51*(2–3), 175–190. http://dx.doi.org/10.1016/j.robot.2004.10.004.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *The Behavioral and Brain Sciences, 30*(1), 63–81. http://dx.doi.org/10.1017/S0140525X07000891 (discussion 81–134).

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain, 120*(4), 701–722. http://dx.doi.org/10.1093/brain/120.4.701.

Neisser, U. (1987). In *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge University Press, pp. x, 317.

Noë, A. (2004). *Action in perception*. MIT Press.

O'Regan, J. K. (2012). How to build a robot that is conscious and feels. *Minds and Machines, 22*(2), 117–136. http://dx.doi.org/10.1007/s11023-012-9279-x.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. New York: Oxford University Press.

Ramachandran, V. S., & Hirstein, W. (1998). The perception of phantom limbs. *Brain, 121*(9), 1603–1630. http://dx.doi.org/10.1093/brain/121.9.1603.

Reichenbach, H. (1958). The philosophy of space and time. *Dover*.

Revonsuo, A. (2006). *Inner presence: Consciousness as a biological phenomenon*. Cambridge, Mass: MIT Press.

Revonsuo, A., & Salmivalli, C. (1995). A content analysis of bizarre elements in dreams. *Dreaming, 5*(3), 169–187. http://dx.doi.org/10.1037/h0094433.

Rockwell, W. T. (2005). *Neither brain nor ghost*. Cambridge, Mass: MIT Press.

Sacks, O. (1970). The man who mistook his wife for a hat: And other clinical tales. *Berkeley*.

Schaffer, J. (2014). The metaphysics of causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/archives/sum2014/entries/causation-metaphysics/>.

Schwitzgebel, E., Huang, C., & Zhou, Y. (2006). Do we dream in color? Cultural variations and skepticism. *Dreaming, 16*(1), 36–42. http://dx.doi.org/10.1037/1053-0797.16.1.36.

Seth, A. K., Baars, B. J., & Edelman, D. B. (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition, 14*(1), 119–139. http://dx.doi.org/10.1016/j.concog.2004.08.006.

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and consciousness in the space of possible minds*. USA: Oxford University Press.

Shanahan, M., & Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition, 98*(2), 157–176. http://dx.doi.org/10.1016/j.cognition.2004.11.007.

Sharma, J., Angelucci, A., & Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature, 404*(6780), 841–847. http://dx.doi.org/10.1038/35009043.

Sharma, J., Dragoi, V., Tenenbaum, J. B., Miller, E. K., & Sur, M. (2003). V1 neurons signal acquisition of an internal representation of stimulus location. *Science, 300*(5626), 1758–1763. http://dx.doi.org/10.1126/science.1081721.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074.

Sur, M., Angelucci, A., & Sharma, J. (1999). Rewiring cortex: The role of patterned activity in development and plasticity of neocortical circuits. *Journal of Neurobiology, 41*(1), 33–43.

Sur, M., Garraghty, P. E., & Roe, A. W. (1988). Experimentally induced visual projections into auditory thalamus and cortex. *Science, 242*(4884), 1437–1441.

Sur, M., & Rubenstein, J. L. R. (2005). Patterning and plasticity of the cerebral cortex. *Science, 310*(5749), 805–810. http://dx.doi.org/10.1126/science.1112070.

Thelen, E. (2000). Grounded in the world: Developmental origins of the embodied mind. *Infancy, 1*(1), 3–28. http://dx.doi.org/10.1207/S15327078IN0101_02.

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology, 63*, 483–509. http://dx.doi.org/10.1146/annurev-psych-120710-100412.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(42), 1–22. http://dx.doi.org/10.1186/1471-2202-5-42.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin, 215*(3), 216–242.

Tononi, G., Edelman, G. M., & Sporns, O. (1998). Complexity and coherency: Integrating information in the brain. *Trends in Cognitive Sciences, 2*(12), 474–484. http://dx.doi.org/10.1016/S1364-6613(98)01259-5.

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Sciences, 1124*, 239–261. http://dx.doi.org/10.1196/annals.1440.004.

Yablo, S. (2004). Advertisement for a sketch of an outline of a prototheory of causation. In N. Hall, L. A. Paul, & J. Collins (Eds.), *Causation and counterfactuals* (pp. 119–137). Cambridge, MA, USA: MIT Press.

Yourgrau, P. (2005). *A world without time: The forgotten legacy of Godel and Einstein*. New York: Penguin Books.

Zeki, S., & Bartels, A. (1999). Toward a theory of visual consciousness. *Consciousness and Cognition, 8*(2), 225–259. http://dx.doi.org/10.1006/ccog.1999.0390.