International Journal of Machine Consciousness
Vol. 1, No. 1 (2009) 1–9

World Scientific
www.worldscientific.com

THE POTENTIAL IMPACT OF MACHINE CONSCIOUSNESS IN SCIENCE AND ENGINEERING

IGOR ALEKSANDER

Electrical and Electronic Engineering Department,
Imperial College, London, UK
i.aleksander@imperial.ac.uk

This paper critically tracks the impact of the development of the machine consciousness paradigm from the incredulity of the 1990s to the structuring of the turn of this century, and the consolidation of the present time which forms the basis for guessing what might happen in the future. The underlying question is how this development may have changed our understanding of consciousness and whether an artificial version of the concept contributes to the improvement of computational machinery and robots. The paper includes some suggestions for research that might be profitable and others that may not be.

Keywords: Machine consciousness; phenomenology; conscious robots.

1. The Incredulous Days

While computer scientists believed that they had a good grip over Artificial Intelligence, they were wary of suggesting that this had anything to do with consciousness. The well known attack on computational approaches to consciousness was launched by Roger Penrose in 1989.[1] This was based on the idea that the insight humans have in the solution of, say, a mathematical problem is not part of their mathematical skill, but seems to enter their consciousness from out of nowhere. He concluded that this phenomenon was outside the realm of logic and mathematics, rendering any talk of a computational approach to consciousness futile. Marvin Minsky in 1991[2] argued against this objection by pointing out that mathematical insight is merely an extension of human ability to internalise experience and not, as Penrose presented, a non-analytic relationship between brain and mind. Of course, at the time, with the exception of machine learning paradigms, computing ways of internalizing experience were not greatly explored. As consequence, in 1996, Chalmers' influential thesis[3] dubbed the mind/body relationship as the "hard problem" of explaining the "supervenience" of sensation on a physical substrate. This allowed paper reviewers to reject submissions which held that consciousness could be studied through computational means. To detect whether a system was conscious or not could no more be

identified in artificial systems than it could in living ones, leaving the designer bereft of ways of creating conscious systems. Chalmers had skilfully raised the idea of a Zombie to underline that behavior is no indicator of the presence of consciousness and computational methods only addressed functional issues of behavior. This made it hard to publish papers on computational approaches to consciousness — referees would reject on the basis of Chalmers' arguments.

During this period, however, there began to appear a few green shoots of the way the above objections may be set aside and that computational methods could be beneficial in forming a science of consciousness. The present author found that a 1996 technical book[4] with "consciousness" admittedly in the second line of the title, was happily received by computer scientist reviewers. For example, William Clocksin of the Cambridge Computer Laboratory[5] showed appreciation of the idea that the computational theory needed for the modeling of consciousness was that of neural automata which, through a process of "iconic" transfer gave the states of the automaton an intentional character.

Also, in 2000, the author published a paper that contained the words "visual consciousness" in the title[6] in the Proceedings of the Royal Society. The paper provided a computationally verified theory of the way that the muscular action of the eye can be responsible for providing a single sensation for disparate activity in the brain relating to a single object in the "out-there-world". Later this was to become the basis for axiom 1 in an axiomatic (that is, initial and not the result of a proof) definition of the components of consciousness.[7]

In addition to the above, there were many detractors to the idea of approaching consciousness from a computational standpoint. Without direct references, the names of Greenfield, Tallis, Lucas and Velmans come to mind as sceptics. It is not suggested here that this scepticism has gone away, merely that the machine consciousness paradigm is gaining confidence.

2. The Formational Days

As is often said, the single event that gave legitimacy to asking whether a machine could be conscious was the Swartz Foundation meeting on this very topic in May 2001 at the Cold Spring Harbor Laboratories.[8] Organized by Christof Koch, David Chalmers, Rod Goodman and Owen Holland, the meeting's conclusion by Koch is also often quoted:

> "*The only* (*near*) *universal consensus at the workshop was that*, *in principle*, *one day computers or robots could be conscious. In other words*, *that we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artefacts designed or evolved by humans.*"

The scientific effect of this was to counter objections that life is needed for consciousness and encourages work in the artificial domain. While this does not suggest

that methodologies for designing conscious artefacts are known in an agreed fashion, it provides a framework for further development of other work presented at the meeting. Then, looking at some detailed contributions it is clear that some important views were presented which established lines of research for a few years ahead. (Note that in the paragraphs referring to presentations at the workshop the common reference is 8 unless others need to be quoted.)

It was clear that Baars' Global Workspace model[9] was taken by many to be the salient computational way of describing a conscious process. It involves the competition of the states of sub-processes of which one wins, enters an architectural area called Global Workspace (GW), from which the winning state is broadcast throughout the system. This controls further developing states in the competing processes and the way they react to incoming information. The sequence of states in the GW is taken to be "the stream of consciousness" of the system. Stan Franklin presented an implementation (IDA: Intelligent Distribution Agent) of the Baars model which was directed at finding billets for US Navy personnel. The benefit of the presence of "consciousness" is that the system might be construed as being "caring" by its users who would normally have communicated with a human. However, this immediately gives rise to the problem that behavior is a poor indicator of the presence of consciousness. Franklin asks whether IDA is "sentient" but concludes in the negative. However, he is confident that sentience can be "added". This has proved not easy and requires a firm outlook on mechanisms for phenomenal states. Franklin returns to this question in 2008, as will be seen in Sec. 4. The Global Workspace model also features in the presentation by Stanislas Dehaene who, with Lionel Naccache,[10] used it to study the cerebral mechanisms of masked priming effects in the brain. This is a good example of the way that a computational model derived within cognitive science may be directed back towards explaining experimental results found in neurophysiology.

Another long-lasting notion was presented by Giulio Tononi who (originally with Gerald Edelmann[11]) argues that only certain neural networks have the power of "integrating" information sufficiently to match such integration that appears to be present in areas of the brain. It should be explained that integration is a measure of the ability of a localized area of the brain to have a rich state structure with states that support the sentient phenomena that are reported by the conscious organism. This work still stirs interest in machine consciousness laboratories and is in need of being taken further forward.

Other contributions included that of the present author who laid down the basic arguments that later would become known as the "axioms" of consciousness. An important point made here was that a conscious robot would have to be conscious of being a robot and not a human. Susan Blackmore advocated that machine consciousness be based on "memes", that is, idea replicators due to Richard Dawkins. This notion, while appealing, did not have the rigorous background necessary for it to contribute to the machine consciousness paradigm of the present day. Perhaps some

day it will be revived. Chris Frith drew attention to the importance of the "other mind" element of consciousness. This too could do with being revisited in contemporary work. A significant presentation on a simulated robot was that by Holland and Goodman who argued that the basis of consciousness in such systems must lie in autonomy and a simulation of itself in a dynamic world. This led to the most important work by Holland to be reviewed below.

Those present, even if not presenting a shared or dominant view, felt at a final discussion, that a compendium of approaches developed from algorithms, neural systems, control systems and analysis would benefit both the definitional and applicational characteristics of machine consciousness. A new paradigm had apparently been born.

3. The Consolidation Days

It is argued in this paper that the common features of the consolidation, which has taken place in the years that have intervened since 2001, have been found mainly in *decomposition* and *internal simulation for robots.* It should be mentioned immediately that some important contributors to machine consciousness were not present at the CSHL meeting but made their impact in these intervening years.

3.1. *Decomposition*

In 2003, Pentti Haikonen published a clear-minded book treating conscious systems from a point-of-view design in the neural engineering domain.[12] His key argument is that one needs to model consciousness as several phenomena, which despite their separation, can act as an interacting whole. He includes perception, motor function, emotion, inner speech, thought and imagery as the separate elements that require modeling, both individually and as an integrated complex. At about the same time, the present author[7] arrived at a very similar procedure of decomposing consciousness into what appeared to be essential features which, on account of their discovery through introspection, he called axioms. There were five of these: presence in an out-there world; imagination and recall of past presence experience; attention as a determiner of the content of experience; volition and emotion. The key point made then was that the emergent dynamic properties of a neural network can be harnessed (through a transfer of sensory information into state coding) to represent experience of the world as a state structure. This exists to provide imaginative experiences and as the basis of volitional planning validated by "emotions". Hence, in contrast with Haikonen, the axiomatic mechanisms were modeled directly as different aspects of the properties of a dynamic neural mechanism.

Also at the same time, German philosopher Thomas Metzinger[13] suggested that conscious machines were possible, but could not be achieved due to the lack of resolution offered by computational techniques known to date. He too felt the need for decomposition which he expressed as "constraints" on consciousness. Many of

these, including the "global availability" and a "window of presence" are a response to the same concerns that led to Aleksander's axiom 1: a phenomenal state needs to be physically present and causal in many of the sentient and behavioral characteristics of a model. From the position of this review, Metzinger's coarseness objection does not prevent neural experimentation to validate the possible operation of decomposed models as has been indicated in the more recent publications by Haikonen[14] and Aleksander[15] where even quite modestly sized neural networks have been shown to have sufficiently rich state structures to support the physical/mental relationships advocated by the authors.

Not all attempts at decomposition have been carried out at a neural level. An influential paper by Aaron Sloman and Ron Chrisley[16] argued for virtual designs of conscious machines. This means that virtual machines should reflect a designer's theory of what it is for a machine to be conscious. Actual implementation can then be achieved in several ways. As an example, the authors present a scheme that has three vertical layers — one for sensory processing, a second for recognition and a third for action — which interact with three horizontal layers. These control direct reaction between input and action at the lowest level, a deliberative link (that accesses stored experience) at the middle level and a managerial layer that monitors the success of the organism and adjusts its operational parameters to improve performance. Alarm mechanisms are also discussed in this schema. Here too, under the heading of "virtual machine functionalism" the authors not only advocate decomposition but argue that the discussion of decomposed models can take place without reference to physical implementation. This is an example of the way that computational ideas (virtualism in this case) can add to philosophical discourse (the mind to body relation in this case).

Finally, it is important to recognize the contribution by Benjamin Kuipers[17] who suggests that the process of consciousness is one of attentional tracking of meaning in a flood of input to the system. He calls this "Drinking from the Fountain of Experience". He too believes in decomposition and accepts the 11 features set out by John Searle.[18] In common with the decompositions used by others in this section of the paper, Searle believes that such features are "quantitative" which can here be interpreted as possibly leading to a computational implementation. Certainly Kuipers shows that this quantitativeness is compatible with his attention-tracking model.

3.2. *Internal simulation for conscious robots*

Another person not present at the CSHL meeting was Germund Hesslow. The focal point of his contribution has been called *the simulation hypothesis.*[19] This requires that the effect of sensory input be stimulated internally and independently of actual sensory input, as should be the actions that might arise. The action representations are vetoed so that they do not actually reach the muscles. In a way similar to Aleksander's axiom 4 (Refs. 7 and 15) the system can select appropriate actions by

"imagining" alternative contingencies in order to select appropriate action. Hesslow and Jirenhed[20] showed that the simulation model could be learned by an on-board neural network so that the robot could *imagine* its previous journeys in an environment that, for example, contained corridors. That is, the robot may be considered as creating an inner world that resembles its environment.

Owen Holland[21] points out that psychologist Kenneth Craik had suggested, as early as 1943,[22] that an organism that carried in its head a model of reality and itself in it, could behave intelligently in its world. Holland was the first researcher in the UK to receive funding for work on conscious robots. He made two salient contributions. The first was to advocate the development of consciousness alongside an autonomic system responsible for embodying the robot in its environment. To this end he built a torso robot that contained some of the skeletal and muscular features of an organism with a spine and limbs. The second was to embrace the internal simulation ideas suggested by Craik[22] and Hesslow,[19] but to do so in stages. These include a simulation of self and environment, the ability of self's knowledge of self and environment and the outcomes of interaction of self and environment. Only as a result of these purely internal activities could a scheme for deliberative action in the environment be developed.

Finally, in the area of robots and consciousness, the work of Antonio Chella needs to be acknowledged.[23] To ground his work, Chella developed a robot that could find its way around the Archeological Museum of Agrigento in Sicily. The key to the consciousness of this robot is that it anticipates in full visual detail the results of possible actions in the environment. The theoretical issue is that the internally anticipated scenes can be recognised in the sense that they can drive the actuators of the system to move between the actual current position to a desired anticipated one.

4. Current Concerns

It has been argued above that two major anchors for doing work in machine consciousness were laid down: decomposition and internal simulation (in the context of robotics). Also a brief allusion was made to the importance of virtualism as a way of escaping from tight philosophies of physicalism and notions of information integration. However, it has become increasingly the case that researchers have included phenomenal states in their designs. Therefore it is important here to review the status of specific concerns in *phenomenology* and those other issues which are treated in machine consciousness laboratories at the moment.

4.1. *Phenomenology*

Phenomenology is a study of consciousness said to have been founded by German philosopher Edmund Husserl who defined it in 1901 as: "The reflective study of the essence of consciousness as experienced from the first-person point of view"[24]. Indeed, in a recent thesis on machine consciousness, David Gamez has defined

consciousness as "The presence of a phenomenal world".[25] A phenomenal machine system could therefore be defined as one which is studied through a concern for internal state(s) which have a capacity for representing reality directly. How could such phenomenal worlds actually arise in artificial systems? In 2007, Aleksander and Morton[26] gave an answer to this by arguing that a neural state machine can "internalize" the world into the state of a neural state machine, which is then used in their axiomatic architecture cited above.

At the same time, Chrisley and Parthmore[27] suggested that a distinction should be created between synthetic phenomenality and synthetic phenomenology. The first indicates the presence of a depictive state as in Ref. 26, while synthetic phenomenology is the effort of describing to the external world the composition of a phenomenal state. In Ref. 26, Aleksander and Morton made the entire state available for inspection on a computer screen, while in Ref. 27, Chrisley and Parthmore sought ways of decoding the non-conceptual content of visual experience.

Even the Global Workspace model has been considered for the possibility of possessing phenomenal states. Stan Franklin, Bernard Baars and Uma Ramamoorty have suggested that mechanisms that generate coherent subjective states can be added to the GW architecture.[28] The present author believes that such states cannot just be added, but should be the focus around which the supporting architecture evolves. The idea that a brain box on top of other brain machinery is responsible for the phenomenal sensation sounds erroneous.

4.2. *Other concerns*

Being a young paradigm there are still a variety of ideas that are in different stages of being established. Here are some samples chosen in an entirely non-exhaustive way. *Attention* is undoubtedly an important aspect of consciousness. John Taylor[29] uses a control model of attention to discover the point at which the mechanization of consciousness can begin. Ricardo Sanz[30] argues that consciousness and self-awareness are necessary features needed to control ultra complex systems. Interesting philosophical points of view are also present among current concerns. Philosopher Susan Stuart[31] supports the decomposition stance seen earlier in this paper, particularly the necessity for embodiment and points out that the roots of such explanatory beliefs may be found in the philosophy of Kant. Another philosophical input comes from Riccardo Manzotti.[32] Under the heading of "externalism" he develops the idea that internal thought needs to be considered as being integrated with those external processes of which the organism is said to be conscious as part of the interlinked and external processes.

Attempts to differentiate networks capable of sustaining consciousness from those that cannot has remained a matter of important theoretical concern. Giulio Tononi (with Christoph Koch) continues to contribute on information integration[33] while, in a similar vein, Anil Seth[34] is pursuing ideas of causality in networks, which is an

alternative to tracking information processing and coding of network signals based on network structure.

5. Science, Engineering and the Task Ahead

In referring to science and engineering in the title of this paper, attention is drawn to alternatives to the classical approaches to the science of consciousness (mainly in the neurosciences). It has been argued that the machine consciousness paradigm has grown from a realization that engineered objects might have subjective feelings. The task is to discover constructivist methodologies that would capture subjective feelings. It is in this quest that the need to *decompose* consciousness into elements for which constructivist procedures might be found has emerged. Interestingly, this has been concluded almost independently by several researchers. Many of these have adopted the stance of dynamic machines with states that have a phenomenal type of reflection of the reality in which such artificial organisms are sited. It is also noteworthy that despite the possibility of discussing consciousness at a virtual machine level, computer science contains methods that map virtual models into physical structures. The neural methodology has a certain edge in this endeavor as it has the capacity not only of structuring palpable machines like robots, but also of feeding theoretical ideas back into neurology and advancing the science of living consciousness in this way.

So, is it all sown up? Not a bit of it! Most of what has been discussed in this paper is about setting frameworks within which important questions (both of competence and existence) can be addressed. How valid are machine models? How persuasive are machine models? Are the key philosophical questions being addressed? What are the uses of machine consciousness? These are just a few of the important questions that will be addressed as the topic of machine consciousness continues to develop.

References

1. R. Penrose, *The Emperor's New Mind* (Oxford University Press, Oxford, 1989).
2. M. Minsky, Conscious machines, in *Machinery of Consciousness*, Proc. National Research Council of Canada, 75th Ann. Symp. Science in Society, June 1991.
3. D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, Oxford, 1996).
4. I. Aleksander, *Impossible Minds: My Neurons My Consciousness* (Imperial College Press, London, 1996).
5. Clocksin: Mind of a Machine. Book review, *Nature* **383** (1996) 592.
6. I. Aleksander and B. Dunmall, An extension to the hypothesis of the asynchrony of visual consciousness, *Proc. R. Soc. Lond.* **267** (2000) 197–200.
7. I. Aleksander and B. Dunmall, Axioms and tests for the presence of minimal consciousness in agents, *J. Consci. Stud.* **10** (2003) 7–19.
8. The Swartz Foundation, Can a machine be conscious? `http://www.theswartz-foundation.org/banbury_e.asp`
9. B. Baars, *In the Theater of Consciousness* (Oxford, Oxford University Press, 1997).
10. S. Dehaene and L. Naccache, Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework, *The Cognitive Neuroscience of Consciousness*, **7** (2001) 1–37.

11. G. Tononi and G. M. Edelman, Consciousness and complexity, *Science* **282** (1998) 1846−1851.
12. P. O. Haikonen, *The Cognitive Approach to Conscious Machines* (Imprint Academic, Exeter, 2003).
13. T. Metzinger, *Being No One* (The MIT Press, Cambridge, Massachusetts, 2003).
14. P. O. Haikonen, *Robot Brains: Circuits and Systems for Conscious Machines* (John Wiley and Sons, Chichester, 2007).
15. I. Aleksander, *The World in My Mind, My Mind in the World: Key Mechanisms or Consciousness in Humans, Animals and Machines* (Imprint Academic, Exeter, 2005).
16. A. Sloman and R. Chrisley, Virtual Machines and Consciousness, *J. Consci. Stud.* **10** (2003) 133−172.
17. B. Kuipers, Drinking from the firehose of experience, *Artif. Intell. Med.* **44** (2008) 155−170.
18. J. R. Searle, *Mind: A Brief Introduction* (Oxford University Press, 2004).
19. G. Hesslow, Conscious thought as simulation of behavior and perception, *Trends in Cognitive Sciences* **6** (2002) 242−247.
20. G. Hesslow and D.-A. Jirenhed, The inner world of a simple robot, *J. Consci. Stud.* **14** (2007) 85−96.
21. O. Holland, R. Knight and R. Newcombe, A robot-based approach to machine consciousness, in A. Chella and R. Manzotti, eds. *Artificial Consciousness* (Imprint Academic, Exeter, 2007), pp. 156−173.
22. K. J. W. Craik, *The Nature of Explanation* (Cambridge University Press, 1943).
23. A. Chella, Towards robot conscious perception, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, Exeter: 2007), pp. 124−140.
24. E. Husserl, Logical investigations, Trans. J. N. Findlay (London Routledge, 1973).
25. D. Gamez, *The Development and Analysis of Conscious Machines*, University of Essex, PhD thesis in computing, 2008.
26. I. Aleksander and H. Morton, Depictive architectures for synthetic phenomenology, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, Exeter, 2007), pp. 67−81.
27. R. Chrisley and J. Parthmore, Synthetic Phenomenology, *J. Consci. Stud.* **14** (2007) 44−58.
28. S. Franklin, B. J. Baars and U. Ramamurthy, A phenomenally conscious robot? *APA Newsletters* **8** (2008) 2−4.
29. J. Taylor, Through machine attention to machine consciousness, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, Exeter, 2007), pp. 24−47.
30. R. Sanz, I. Lopez and J. Bermejo-Alonso, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, Exeter, 2007), pp. 141−155.
31. S. A. J. Stuart, Machine consciousness, cognitive and kinaesthetic imagination, *J. Consci. Stud.* **14** (2007) 141−153.
32. R. Manzotti, An alternative process view of conscious perception, *J. Consci. Stud.* **13** (2006) 45−79.
33. C. Koch and G. Tononi, Can machines be conscious? *IEEE Spectrum* **45** (2008) 54−59.
34. A. K. Seth, Causal networks in simulated neural systems, *Cognitive Neurodynamics* **2** (2008) 49−64.