Original Paper UDC 165.12: 167.7: 141.33
Received November 15th, 2007

Igor Aleksander
Imperial College London, Department of Electrical and Electronic Engineering – Exhibition Road, GB-London, SW7 2BT
i.aleksander@ic.ac.uk

Modeling Consciousness in Virtual Computational Machines

Functionalism and Phenomenology*

Abstract

This paper describes the efforts of those who work with informational machines and with informational analyses to provide a basis for understanding consciousness and for speculating on what it would take to make a conscious machine. Some of the origins of these considerations are covered and the contributions of several researchers are reviewed. A distinction is drawn between functional and phenomenological approaches showing how the former lead to algorithmic methods based on conventional programming, while the latter lead to neural network analyses. Attention is drawn to the many open questions that this approach generates and some speculation on what future work might bring is included.

Keywords

science of consciousness, computer Modeling of cognition, phenomenology in machines, virtual machines

1. Introduction: Origins

Undoubtedly, a highly influential discourse on consciousness has been that of David Chalmers (1996) who suggested that explanations of consciousness based on the grounding of conscious experience in the mechanics of the brain (neurophysiology and neuroanatomy) only addresses the "easy" problem of what is necessary to have cognition (memory, attention, language etc.). He argues that conventional science leaves the "hard" problem of what it is to have conscious sensations (sometimes, *qualia*) untouched. However, in the latter parts of his discussions he speculates on the distinctiveness of science based on *information* (as opposed to physics, chemistry or biology) as a possible approach to solving the hard problem.

So, machine consciousness is an attempt to understand consciousness using the methods and laws of informational machines. Such machines, on the one hand, depend on specific programs (or algorithms) so designed that external behaviour leads an observer to attribute consciousness to the machine. On the

*
This is an elaborated version of "Machine Consciousness" submitted by the author to www.scholarpedia.org.

other hand there are other machines which lead to fine-grain neural systems where with internal states that have sensation-like characteristics.

An important event in the history of this topic is a meeting sponsored in 2001 at the Cold Spring Harbour Laboratories (CSHL). Sponsored by the Swartz Foundation (that normally funds scientific meetings on brain studies) it addressed the question 'Could Machines Be Conscious?'. The organisers were neuroscientist Christof Koch, philosopher David Chalmers and computer scientist, Rodney Goodman. While there was little agreement on precise definitions of consciousness between the audience of 21, made up of neuroscientists, philosophers and computer scientists, there was agreement on the following proposition. "There is no known law of nature that forbids the existence of subjective feelings in artefacts designed or evolved by humans" (http://www.swartzneuro.org/abstracts/2001/summary.asp). In the years which followed several lines of research investigating the design of such artifacts, came into being and these form the basis of this article. We also consider some of the approaches that were discussed at the CSHL meeting on which the above conclusion was reached.

2. Early Models

2.1 Global Workspace

One of the models of conscious processes that pre-dates the CSHL meeting is that of Bernard Baars (1988 and 1997). The model is based on the supposition that there are many unconscious processes that compete with one another to cause their output to enter a 'global workspace' (GW) area of the system. Such unconscious processes could be perceptual or memory-based. The GW 'broadcasts' the winning entry to all the unconscious processes influencing their subsequent states. It is this broadcast that, according to Baars, constitutes the conscious event. The entry into the GW is also influenced by the current sensory inputs to the system so that the most important or salient process with respect to current sensory state of the organism enters the GW. This approach has influenced computer scientists to attempt to build sytems to which the user may attribute consciousness.

For example, Stan Franklin of Memphis University (2006) made use of Baars' Global Workspace (GW) model to design a billeting system (a job-placement system) for sailors. The needs of an individual sailor are the competing processes and suggestions for billeting from which the result of the broadcast from the GW constitutes the advice given to the user. Franklin explicitly does not claim that this process encompasses an explanation of sensations, it mainly stresses that functionally, consciousness may be attributed to his working model as, to a user, it appears to carry out a task that normally requires conscious deliberation in a human billeter.

Others influenced by this approach are Murray Shanahan of Imperial College, London (2007) and Stanislas Dehaene et. al. (2003). Shanahan has designed spiking neuron versions of the global workspace notion that enable a robot to carry out 'internally' simulations and predictions of the results of its own actions in the world. "Spiking" is the process whereby neurons in the brain communicate, and use of such techniques may apply directly to an understanding of the way that the brain sustains conscious states. Dehaene and his colleagues have also used a neural network model to throw light on the way

that visual stimuli enter consciousness by entering a global workspace and being broadcast to distant areas.

2.2 Virtual Machine Functionalism

A virtual machine is an entity that 'runs' on a computer with characteristics of its own that are independent largely of the processing nature of the 'host' computer. For example, most computers have a virtual 'calculator' button which causes the face of a calculator to be displayed on the screen. By means of a 'mouse', the buttons on this calculator can be pressed as if they were on a real, physical, calculator achieving the same results as a physical calculator. Now the physical makeup of a real calculator and a virtual one are quite different. But the ability to calculate for the two can be discussed in precisely the same terms. Philosophically this is important as it breaks the link between the function of a system and the physical substrate that causes it. So, in terms of the 'hard problem' of consciousness, it allows conscious sensation to be discussed using informational techniques on the understanding that a wide variety, possibly an infinity of physical substrates could cause it.

Boston Philosopher Daniel Dennett recognized this by noting that virtuality is a way of resolving over-constraining materialist ties in the study of consciousness as, in a virtual domain it is possible to consider mental states as being the result of a virtual machine running on the parallel material of the neural brain. This does not depend on the way in which the material of the brain *causes* the mental state. Sloman and Chrisley (2003) developed an approach to machine consciousness which develops Dennett's suggestion towards what they call "virtual machine functionalism". They draw attention to Block's (1996) contention that in a functionalist stance, a mental state may be likened to the state of a state machine which only leads to a new state in function of a new sensory input. Sloman and Chrisley point out that this impoverished view may be replaced by a richer one where one allows that a mental state is the product of the state of several interacting machines whose states not only contribute to an overall mental state, but where these sub-states can modify each other through interaction of the machines. Calling this 'virtual machine functionalism', they provide a 'schema' (COGAFF) for discussing the virtual processes that constitute a functional view of being conscious. They suggest that it is important to consider two principal interacting streams, one which flows from sensation to action and another which flows from reaction to deliberation. They also allow for a global form of altered control such as may arise in an emergency.

2.3 Phenomenal Virtual Machines

A partitioned mental state as advocated under virtual machine functionalism also appears in Aleksander (2005). Here, some of the sub-states have a phenomenal character through being assumed to be the states of fine-grained neural networks which 'depict' the world and the organism in it. In implementations, such depictions are displayed on computer screens to make explicit the virtual mental state of the machine. To shed some computational light on consciousness, Aleksander argues that it is a concept that covers many phenomena. He expresses these as five axioms for each of which meaningful virtual states of neural computing models may be found. The first is 'presence' which finds mechanisms for representing the world with the organism in it

by recognizing that motor signals (such as resulting from eye movement and head movement) need to influence the representation of information coming purely from the sensing surfaces such as eyes, hands, etc.. This results in an inner 'depiction' of the external world. The second is 'imagination' which relates to our ability to visualize the world in the absence of sensory input. This may be explained by evoking results in neural network theory which show that depictions can be sustained as a result of re-entrant (feedback) connections within the network. The third is 'attention' which (exogenously) relates to mechanisms that guide the sensors of the organism during perceptual acts and (endogenously) when such guidance occurs during imaginative acts. The fourth and fifth axioms deal with the related concepts of 'planning' (exploration in imaginative acts in the presence of an inner volitional state) and 'emotion' as an evaluation of plans. For example, in a restaurant a 'volitional' state may be determined by the liking (parallel emotional state in a neural net) of some offerings rather than others and weighing up other factors such as a fear (other parallel emotional state) of putting on weight.

3. Consciousness in Systems and Robots

3.1 Conscious Processes as Algorithms or Control Structure

With advancing technology, engineers are facing a worrying increase in the complexity of the systems they are designing. Through this they notice that they may have something to add to the understanding of how the human brain, the most complex system on earth, might achieve its major feat: the sustaining of consciousness. Designers of complex systems both in artificial intelligence (e.g. chess playing) and control structures (e.g. the control electronics for a jumbo jet) have pointed to aspects of such design as being helpful in modelling and understanding consciousness. For example, Benjamin Kuipers of the University of Texas (1996) draws attention to algorithms in artificial intelligence which extract and track specific items of information from what he calls "the firehose of experience". It is the very process of tracking, possibly a multiplicity of objects and events which endows the machine with the property of generating a coherent narrative of the world. In humans, this would be called giving an account of that of which it is conscious. While admittedly addressing the Chalmers' 'easy problem', Kuipers suggests a way of "sneaking up" on the 'hard problem' citing computational systems to hold representations which correspond to world events and processes by being created through algorithms that *learn* such representations. This admittedly appears to leave open the question as to why any informational constructs should be *felt*. Kuipers suggests that whatever "feeling the information" might mean, it is likely to be derived from reasonably accurate representations of the world and the ability to act correctly, these being the direct product of appropriate informational processes. In a similar vein, but in the context of control system design, Ricardo Sanz of Madrid University (2007) points out that the machinery which is expected to behave correctly in a complex world demands a design complexity that relies increasingly on adaptation and learning. He suggests that there may be something it is like "to be a model-based reflective predictive controller" (a very advanced form of control system) of a machine with a mission, which is akin to there being something it is like to be a conscious being with a purpose in life.

3.2 Consciousness for Robots

Owen Holland of Essex University with his colleagues (2007) has the distinction of having been the first researcher to have been awarded a major grant [by the Engineering and Physical Sciences Research Council in the UK] to investigate the basic constructionist question: 'were a robot to be conscious, how would it be designed?' Holland's primary principle is to build a human-like skeletal structure ready to engage with a real world so that it can build an internal virtual model of the world and its own interaction with it as a fundamental conscious thought. This internal model is based on Gerry Hesslow's notion of consciousness as 'inner simulation' developed at the University of Lund (2007). This too has led to the concept-proving design, with Tom Ziemke, of a miniature robot called K. Part of Hesslow and Ziemke's philosophy is not to ask whether the artifact has or has not qualia, but to note that the question can be asked of the robot with the same legitimacy as is done for humans or other animals.

3.3 The Externalist Outlook for Robot

Antonio Chella of Palermo University (2007) has designed a robot that accepts two views of consciousness: externalism (as advocated by Riccardo Manzotti (2006) of the University of Milan) and the 'sensorimotor contingency' expressed by Kevin O'Regan and Alva Noë (2001). Externalism contends that consciousness of the world cannot be studied by constraining it to the brain – if a theory of consciousness is to arise it must incorporate the 'entanglement' of brain and environment. In agreement with this is the sensorimotor contingency which maintains that there is little need for internal representation in an organism conscious of its environment – the world as it stands provides sufficient representation. The motor parts of the organism are crucial, and 'mastery' of the sensorimotor contingency suggests the learning of appropriate action which attends to important features in the world. Consciousness is then a 'breaking into' this process. Chella has designed robots that are intended to act as guides to human visitors to a museum. The two theories impinge on this design because there is a tight coupling between robot and environment as the robot has mastery of a sensorimotoy contingency to guide its action.

4. Neural Mechanisms

4.1 Neural Models of Cognition

Looking at systems other than robots, a substantial contribution to machine consciousness has been made by the Finnish engineer Pentti Haikonen. Working for the Nokia company in Helsinki he is developing a neural-network approach to express important features of consciousness as signal processes in a multi-channel architecture (2003). He addresses issues such as perception, inner vision, inner speech and emotions. In common with others who approach consciousness as an engineering design process, Haikonen comes to the conclusion that given a proper decomposition of the concept into its cognitive parts, many mysteries are clarified. In his recent writings (2007) Haikonen confirms and deepens the notion that his neural structure leads to an informed discussion about meaning and representation. Of course, with the engineer-

ing approach, the criticism can be advanced that all this only addresses the 'easy problem' of the necessary functioning of substrates to create cognitive representations leaving the 'hard problem' of the link to experienced qualia or sensations untouched. The main counter to this argument is currently the 'virtual' notion that qualia and sensations may be discussed as virtual concepts that do not depend on links to physical substrates and yet can be addressed in the language of informational systems.

4.2 Theory within Neurophysiology

While discovering behaviours of the cells (neurons) in the brain itself is not a direct contribution to machine consciousness as discussed in this article, it still plays an important role both as inspiration for design and to identify what still needs to be discovered in the brain in order for convincing machine models to be developed. A typical example of the former is the work of John Taylor of King's College in London. He argues for the central role of attention in human consciousness, and that, consequently, machine approaches should deal centrally with attention. In Taylor (2007) he describes a system called CODAM (Corollary Discharge of Attention Movement) and argues that 'bridging' links may be found between the model and phenomenal consciousness in humans, paricularly in terms of transparency, presence, unity, intentionality and perspective. A deeper theoretical approach is suggested by Seth et. al. (2006) that draws on the work of Tononi (2004). This introduces a measure of 'Information Integration', (Φ) which treats consciousness as a necessary capacity of a neural network rather than a process. Relevant in these assessments is a measure of causally significant connections in neural networks that can be seen as repositories of experience (Seth 2005). In sum, this informational analysis related to consciousness looks for a multiplicity of measurements which, when taken together, identifies a level of complexity in a neural network that may be needed to achieve the complexity of organisms which are likely to be conscious.

5. Concluding Perspective

Currently (at the very beginning of 2008), machine consciousness has the status of a pragmatic attempt to use methodologies known in information, neuronal and control sciences both to throw light on what it is to be conscious and to aid the design of complex autonomous systems that require a non-living form of consciousness to act correctly in a complex world. In the process of early development are the notions of both functional and phenomenological virtual machines which encourage informational discussions of consciousness in a way that is not limited by coupling to its physical substrate. Machines appear to benefit from this as the acquisition of the necessary skill to operate in highly complex informational worlds is akin to the conscious action of living organisms in similar worlds.

References

Aleksander, I. (2005) *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in Humans Animals and Machines*, Exeter: Imprint Academic.

Baars, B. (1988) *A Cognitive Theory of Consiousness*, Cambridge: Cambridge University Press.

Baars, B. (1997) *In the Theater of Consciousness: The Workspace of the Mind*, New York: Oxford University Press.

Block, N. (1996) "What is functionalism?" (a revised version of the entry on functionalism in *The Encyclopedia of Philosophy Supplement*), Macmillan.

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

Chella, A. (2007) "Towards Robot Conscious Peception", in Chella and Manzotti (Eds.), *Artificial Consciousness*, Exeter: Imprint Academic.

Dehaene, S., Sergent, C. & Changeux, J.-P. (2003) "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data During Conscious Perception", *Proceedings of the National Academy of Science*, 100 (14), 8520–8525.

Franklin, (2003) "IDA a Conscious Artifact? ", *Journal of Consciousness Studies* (4–5), 47–66.

Haikonen, P. O. (2003) *The Cognitive Approach to Conscious Machines*. UK: Imprint Academic.

Haikonen, P. O. (2007) *Robot Brains: Circuits and Systems for Conscious Machines*, UK: Wiley & Sons.

Hesslow. G. and Jirenhed, D. A. (2007) "Must Machines be Zombies? Internal Simulation as a Mechanism for Machine Consciousness", *Proc AAAI Symp. Machine Consciousness and AI*, Washington.

Holland, O., Knight, R. and Newcombe, R.(2007) "The Role of the Self Process in Embodied Machine Consciousness", in Chella and Manzotti (Eds.), *Artificial Consciousness*, Exeter: Imprint Academic.

Kuipers, B. (2007) "Sneaking Up On the Hard Problem of Consciousness", *Proc AAAI Symp. Machine Consciousness and AI*, Washington.

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

O' Regan, K. and A. Noe (2001) "A sensorimotor account of visual perception and consciousness", *Behavioral and Brain Sciences*, 24 (5).

Seth, A.K., Izhikevich, E.M., Reeke, G.N., & Edelman, G.M. (2006) "Theories and measures of consciousness: An extended framework", *Proc. Nat. Acad. Sci*. USA.103(28): 10799–10804.

Sanz, R., Lopez, I. and Bermejo-Alonso, J. (2007) "A Rationale and Vision for Machine Consciousness in Complex Controllers", in Chella and Manzotti (Eds.), *Artificial Consciousness*, Exeter: Imprint Academic.

Seth, A. K. (2005) "Causal connectivity analysis of evolved neural networks duribehavior", *Network: Computation in Neural Systems*, 16(1): 35–54.

Shanahan, M. A (2007) "Spiking neuron model of cortical broadcast and competition", *Consciousness and Cognition*, doi:10.1016/j.concog.2006.12.005.

Sloman, A. and Chrisley, R.(2003) "Virtual machines and consciousness", *Journal of Consciousness Studies*, 10:4–5 (April/May), 133–72.

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

Igor Alexander

Bewusstseinskreierung bei virtuellen Datenverarbeitungsgeräten

Funktionalismus und Phänomenologie

Zusammenfassung

In diesem Beitrag werden die Anstrengungen von Forschern beschrieben, die sich mit Datenverarbeitungsgeräten und Informationsanalysen beschäftigen, um die Grundvoraussetzungen zu schaffen für ein adäquates Verständnis von Bewusstsein sowie Spekulationen darüber, welche Schritte erforderlich sind, um eine mit einem Bewusstsein ausgestattete Maschine herzustellen. Während die Beiträge einiger Forscher im Einzelnen vorgestellt werden, bleiben andere Urheber erwähnter Spekulationen unerwähnt. Der Verfasser unterscheidet zwischen einem funktionalen und einem phänomenologischen Ansatz. Er zeigt auf, dass der funktionale Ansatz in algorithmischen, auf konventionellen Programmierungsmethoden gründenden Methoden resultiert, der phänomenologische Ansatz wiederum in neuralen Netzwerkanalysen. Sodann widmet sich der Verfasser zahlreichen offenen Fragen, die aus diesem Ansatz hervorgehen, und stellt Überlegungen über die möglichen Ergebnisse zukünftiger Forschungen an.

Schlüsselbegriffe

Bewusstseinswissenschaft, Bewusstseinskreierung, Phänomenologie von Datenverarbeitungsgeräten

Igor Aleksander

Modélisation de la conscience dans les machines informatiques virtuelles

Fonctionnalisme et Phénoménologie

Résumé

L'article décrit les efforts employés par ceux qui travaillent sur des machines et des analyses informationnelles afin de fournir des clés de compréhension de la conscience et des hypothèses sur les moyens nécessaires à la fabrication d'une machine consciente. Le texte mentionne certaines origines de ces considérations et fait le compte rendu des contributions de plusieurs chercheurs. Une distinction est établie entre les approches fonctionnelles et phénoménologiques. Les premières mènent aux méthodes algorithmiques fondées sur la programmation conventionnelle, tandis que les secondes mènent aux analyses du réseau neural. Le texte attire l'attention sur de nombreuses questions ouvertes suscitées par cette approche et comporte des réflexions sur des travaux futurs.

Mots-clés

science de la conscience, modélisation informatique de la cognition, phénoménologie des machines, machines virtuelles