Cogn Comput (2009) 1:22–28
DOI 10.1007/s12559-009-9008-9

Designing Conscious Systems

Igor Aleksander

Published online: 23 January 2009

Abstract This paper reviews computational work that is currently developing under the heading of 'Machine Consciousness' and sets out to provide a guide for those who wish to contribute to this field. First, questions of philosophical concern as to the appropriateness of this activity are raised and discussed. Then some classical designs and computational attitudes are described before arguing that fine-grain neural approaches are needed to provide truly phenomenal representations that stand in relation to the behaviour of a computational organism as subjective mental states stand in relation to the existence of a conscious organism. The paper concludes with an evaluation of the validity and benefits of designing conscious systems.

Keywords Machine consciousness · Phenomenology · Conscious robots · Virtual machines

Introduction

The aims of those who contribute to the 'Machine Consciousness' paradigm are first to clarify what it is for an organism, whether it be human, animal or artefact, to be conscious. Second is the aim to examine the potential for informational machines to *be* conscious and what benefit this might bring to the general area of cognitive computation. A brief consideration is given to the philosophical and cultural implications of these developments as it impinges on deeply held beliefs that being conscious is the prerogative of living organisms and cannot be transferred to the domain of informational machines. The sense in which the inner states of an informational machine can be said to be like 'mental' states is addressed stressing that effort in machine consciousness focuses on ways of creating inner states that can be said to be subjective and, in some cases, phenomenal (i.e., world-related).

Computational attempts at creating such subjective states are reviewed with a brief pointer to work done on robotics. Virtual machines are discussed to the extent that they impact on machine consciousness not only as virtual structures capable of the kind of processing that appears to mimic conscious processing in living organisms, but also as helpful constructs that loosen the problematic overtight relationship between mental state structures and their physical underpinnings as may be found in physicalist philosophy.

A phenomenal state in a system is one that is responsible for the behaviour of the system by reflecting the properties of the real world. The nature of computationally phenomenal states is introduced and a distinction between functional and phenomenal virtual machines is drawn as it is beginning to be clear that claims of subjective conscious states cannot be made without phenomenology. The upshot of phenomenal designs is that they essentially evoke neural computation which enables the creation of internal states that reflect the real world. A particular approach previously taken by the author is included for completeness. This is the 'axiomatic/introspective method', which decomposes the concept of being conscious into elements which have reasonably clear transitions into neural architectures. To conclude the paper, some of the many questions related to the further advance in this field are raised and some answers are suggested.

I. Aleksander (✉)
Electrical and Electronic Engineering Department,
Imperial College, London SW7 2BT, UK
e-mail: i.aleksander@imperial.ac.uk

The Engineers' Consciousness Credo and the Credibility Gap

The Optimism

In 2001, the Swartz brain-science foundation organised a three-discipline (philosophy, computation and neuroscience) workshop on the question of 'could a machine be conscious?'. While there were many disagreements, one area of agreement (as summarised by one of the organisers, Christof Koch)[1] was:

> …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.

This statement carries a streak of optimism as well as a challenge for devising ways in which machines with subjective states could be designed. It gave rise to several projects that attempt to do just this: create machines with subjective states that determine the behaviour of the machine. The description of such efforts is the salient topic in this paper.

The Scepticism

In contrast with the above declaration, there have been several expressions of scepticism that need to be made explicit before proceeding to look at computational strategies in machine consciousness. The objections fall into two major classes: the unassailability of Chalmers 'Hard Problem' [1][2] and Penrose's more general notion that consciousness is outside the realm of what is computable [2]. In this paper, the 'hard problem' is addressed by relating it to what is known of the relationship of physical structure to state structure in automata theory. In particular, this puts the onus on machine consciousness researchers to show how the inner states of a state machine can become subjective.

The second objection is cultural with its roots in Aristotle's notion (in *de Anima*) that matters of logic and mathematics are useful in a separate domain from that of observational biology which is the proper way to address the nature of living organisms including thought in human beings. That this is a cultural objection is evident from the Penrose's conclusion [2] where he contends that consciousness is too important to be 'conjured up' by some computation. The 'importance' can most easily be interpreted as a cultural issue.

All this has changed with the advent of computers, particularly in their ability to support virtual machines. This allows organisms normally belonging to the realm of living things to be studied as a virtual machine, that is, a machine that can be investigated as if it were an organism capable of a virtual life, without it actually being alive. It can even be postulated that consciousness can be virtual on the hardware of the brain. This is discussed later. Now, some typical examples of work done by designers of conscious system are given.

Some Existing Computational Approaches

One of the oldest models developed by Baars [3] is known as 'Global Worksapace Theory'. This assumes that there are several *unconscious* processes often quoted in cognitive science (e.g., various forms of memory, volitional and emotional activities) that compete for entry into an architectural element known as the 'global workspace'. The competition is won by the process that is most salient for the sensory input present at the time. A key step follows: the winning contents of the global workspace are broadcast to the competing processing changing their state. This is the 'moment of consciousness' and it is a sequence of such moments that constitutes the system's 'stream of consciousness'.

While this system has no pretence of phenomenal consciousness (i.e., mechanisms that represent the world in detail—see below), a move towards phenomenology was executed by Shanahan [4] using simulated digital neural networks. Shanahan made the unconscious processes (hence the Global Workspace) contain direct visual data. Does 'global workspace theory' have a meaning in neurophysiology? A positive answer was given by Dehaene and Naccache [5] who showed that areas of the brain that include the prefrontal cortex, the anterior cingulate and related regions, form a global workspace and, according to the model, stand in appropriate relation to distant brain areas that carry unconscious memory processes.

Another noteworthy contributor to machine consciousness is Haikonen who published two major books on the subject [6, 7]. He believes that most characteristics of being conscious can be represented in a repetitive architecture of conventional neural networks.

Virtual Machine Functionalism

Functionalism is a philosophical standpoint that addresses the *behaviour* of an organism in a real world as a result the

[1] http://www.theswartzfoundation.org/abstracts/2001_summary.asp.

[2] This suggests that science can only be done on the physical (body) and only correlations can be found to the subjective (mind). Chalmers has argued that the 'hard problem' for science is that it cannot prove that the physical implies the subjective.

effect of that real world on a mental state. In the general case, philosopher Ned Block [8] has pointed out that a functional model of cognition is merely a state machine, where the mental state changes to keep track of a developing environmental reality without any particular restriction on the coding of such states. He illustrates this by suggesting that if a mental state moves from tranquillity into a state of pain, all this instigates is a propensity to 'say ouch' or have other 'thought states' that are contingent on the pain state. There is no attempt to explain the complexity of such a state or how it encodes deliberation. Calling this an 'atomic' view of functionalism, Sloman and Chrisley [9] pointed out that a lack of clarity sets in if the state of a functional system, where many conscious processes may be ongoing, is represented as a single state. This led them to define virtual machine functionalism (VMF) by stating that a functional mental state as one in which many conscious processes are present simultaneously each with its own state structure. For example, a headache state might be accompanied by thoughts of phoning a doctor, the effect on writing a paper, needing to cook a meal and paying one's bills. That is, it is important to recognise that several state machines may be acting simultaneously each providing an element of an overall mental state. Such automata are highly variable, and their essence is 'virtual' in the brain.

We recall that a virtual machine is one that runs on a host machine and the properties of which can be studied independently, without reference to the operation of the host machine. In his explanation of consciousness, philosopher Dennett [10], evoked a virtual machine approach:

> Human consciousness … can best be understood as the operation of a "*Von Neumannesque*" virtual machine *implemented* in the parallel architecture of the brain that was not designed for any such activities.

The key phrase here is that it may be wrong to look for a design that specially supports the states of a functionally conscious system, but that such a system which evolved in order to cope with the complexities of its environment also runs a virtual consciousness as an added bonus. The real importance of virtuality is that among the neurons of the brain, mental states are largely defined by the environment and that a mental state structure will arise tolerating a considerable amount of latitude in the exact physical structure of the supporting neurology. The reference to a 'Von Neumannesque' machine appears unnecessary. The key issues for VFM are that, whatever it is for the machine to be conscious might be expressed as a virtual machine that reflects the complexity of multiple interacting state machines. As even an infinity of physical structures can support such a VM, the trick is to find some bounding constraints. Sloman and Chrisley have done this by identifying interacting layered schemes: *horizontal* going from the reactive to the deliberative to the managerial and *vertical* going from sensory input to its interpretation ending in a system of action.

Robots

Much machine consciousness is done in connection with robots. This is important as leaving everything to simulation causes the virtual system to be conscious only of other virtual material in the computer. In contrast, a virtual 'mind'[3] of a robot needs to become conscious of the real world in which the robot is situated. A framework for the structure of such minds has been researched by Holland and his colleagues [11] and based on an 'Inner Simulation' model of consciousness due to Hesslow [12]. Holland argues that the internal simulation is built up from a knowledge of being in the world through several steps to an inner simulation of the possible interactions between self and world. Holland found it useful to build an anthropomorphic skeletal robot (called CRONOS) that had the opportunity for sensing inner variables such the state of muscles and positions of body parts. This is ongoing work.

Chella also leads a 'robot consciousness' team which, among other ideas, is developing a robot guide for museums [13]. This is largely based on perceptual 'awareness' in vision where a representation of what is expected (called 'imagination' by the authors) is compared with sensory visual data from the environment in order to lead to action.

Virtual Machine Phenomenology

Phenomenology is a study of consciousness said to have been founded by German philosopher Edmund Husserl who defined it as (1901): "The reflective study of the essence of consciousness as experienced from the first-person point of view" [14]. A phenomenal system therefore is one which is studied through a concern for internal state(s) which have a capacity for representing reality directly in a way that is a decent approximation of the external reality. While 'decent' is not defined, it refers to a sufficiently accurate representation of the environment to form the basis of the behaviour of the organism that will

[3] The term 'mind' needs definition within the virtual consciousness paradigm developed here. If a mental state is the current content of the consciousness of an organism, mind, as the capacity of all possible mental states as organised into a state structure, *is* the state structure of the organism.

not lead to gross errors. Such states must be parts of state structures (i.e., a virtual machine) that represent the behavioural experience of the organism. In order to achieve an unrestricted reflection of reality, a fine-grain representation is implied where the grain is determined by the minimal changes in an external world of which the system is to become conscious.

A Definition of a Weightless Neuron for Use in Phenomenal Systems

The required fine grain has been achieved in previous work though the use of *weightless* digital neurons [15]. One type of weightless neuron maps an n-input binary vector X into a binary variable z which can have value 0, 1 and u, where u represents a random choice between 0 and 1. Learning takes place during a training period when a special binary 'teaching' input line d (desired) of the neuron determines whether X is associated with $z = 0$ or $z = 1$ which is stored in the neuron's lookup table which is normally in state u before training takes place. If during a training sequence, the stored value of 0 or 1 is contradicted, the stored lookup state for the contradicted X reverts to the u state.

As generally defined, the weightless neuron also generalises to the extent that if an unknown input vector X_u is compared to the X_j of (X_j, d_j) pairs on which the neuron was trained, and there is a distinct X_j which is closer than any other to X_u (in Hamming distance, say), then the neuron will respond with the corresponding d_j.

Iconic Transfer and Phenomenal States

Say that a network consists of k neurons, each with n inputs, which is 'connected' to a pattern P that consists of a bits. The connection is made at random. Then, there exists a set of teaching lines $D = \{d_1, d_2 \ldots d_k\}$ which, after a training step, defines the k-bit output pattern Q. Now, if D is connected to pattern P as well, Q learns to be a sampling of P.

Transferring this now to a recursive network in which the n inputs of each neuron not only sample P, but also Q (possible with a defined ratio), Q becomes the state of a neural automaton. We submit that this is a *phenomenal* state as it depends on P alone which is the interface where the reality of the automaton's environment is represented. Note that the learned states of Q can be sustained when P changes to unknown states which is the basis of the experiential memory in the system. Figure 1 shows the development of a phenomenal state in a 144 × 144 (the dimension of Q) neuron network with a 144 × 144 input (the dimension of P). Each neuron has 288 binary inputs, 144 randomly drawn from the input P and 144 randomly drawn from state Q. This is a model of the tool-making ability of 'Betty', a crow studied in the zoology department at Oxford University.[4]

This weightless neural state machine was trained by being exposed to the shown sequence, illustrating that *iconic transfer* may be used to create state a state sequence that represents past sensory experience. This may be triggered by an initial input state, and the internal sequence then becomes an imaginational representation of future action. When executed, the action leads to the new input state in the lower group which leads to a different internal sequence—one for taking no action.

The reason for referring to this as a quasi-phenomenal representation lies in the fact that it is a 'third person' view and does not attempt to explain the first person experience. To go beyond the third person we briefly look at some previously published introspective axioms [16] and comment on the mechanisms these imply.

Five Axioms

These five axioms are a breakdown of what important elements of consciousness *feel like* and how they may be translated into neural mechanisms: presence, imagination, attention, volition and emotion. The first is explored in some depth and the others cursorily.

Presence: I feel that I am Centred in an out-There World

To achieve this, the 'out-there-world' needs to be phenomenally represented as being unaffected by the actions (e.g., eye movement, head movement, body movement…) of the organism. That is, it makes it possible to represent the independence of the 'self' in the world. To achieve this it is required that whatever sensory input is being represented, is must be compensated for the acquisition actions of the organism. Say that the eye is foveally fixated on the nose of a face. Say we give the position of the nose the vertical plane spatial origin x, y coordinates 0, 0, and allow that an internal phenomenal representation of the nose in a neural area indexed 0, 0. Now say that the gaze shifts slightly to see the ear at coordinates 1, 0 (in non-defined units). This means that a new neural area centred on 1,0 has to be iconically activated. The implication for modelling is that the neural network training of weightless neurons needs to be *indexed* on muscular activity. In vision this causes, eye movements to create a phenomenal inner state larger than the foveal area. There is an interplay between the creation of such immediate phenomenal states that are

[4] http://users.co.ac.uk/∼kgroup/index.html.

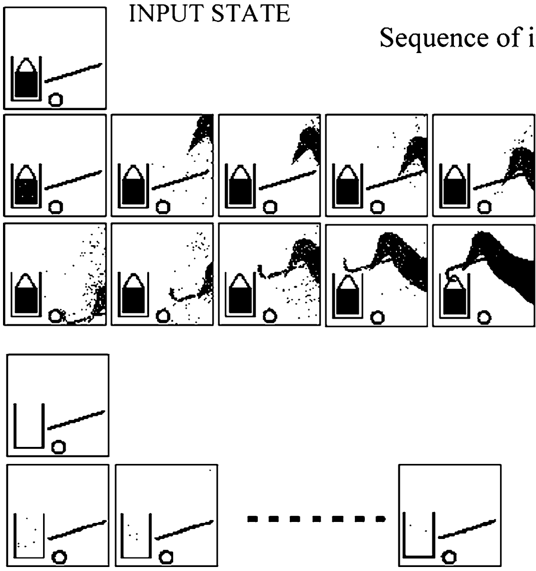

Fig. 1 Quasi-phenomenal behaviour of a 144 × 144 (2076) weightless neural network modelling a crow making the tool that can extract a food canister from a jar. The 2076-neuron network was iconically trained to show that the input state depicting a food canister in a jar and a bendable rod can lead to a sequence of internal states that recalls the way that the problem is solved from experience of a previous successful trial. Each state consists of 144 × 144 binary picture points (outputs of neurons each of which sense 72 other neuron outputs at random and 72 points from the input). Also it is shown that if the jar becomes empty the case of no action can be represented in the internal states

sensed as a forward view and the way that these become parts of a state structure caused by other major movements. That is, a head movement from coordinates x, y to x', y' will cause a related state change without changing the set of neural state variables. In parenthesis, such indexing is highly present in the brain.

Imagination: I Can Remember Not Only Past Experience, But Also I Can Imagine Fictitious Experience

State structures that represent experience can endure without sensory input as a result of the generalising properties of the weightless neurons. That is, the input to a neuron from Q can be sufficient to sustain appropriate state sequences in Q even if inputs from P do not have the values on which the system was trained. Fictional imagination can be shown to be a by-product of language (e.g., an igloo is made of ice bricks)—or random transitions.

Attention: I Am Only Conscious of That to Which I Attend

There are many attention mechanisms in a living organism. These range from unconscious saccades of the eye to salient regions of input (as mediated by the *superior colliculus* in the brain) to purposeful bodily movements that (say) reveal hidden parts of the visual environment.

Volition: I Can Select What I Want and Can Act to Obtain It

State structure can be transversed under various control schemes (e.g., hunger causes food source trajectories to be traversed). This is a form of planning as it is controlled by a 'need' state in part of the automaton. It works together with the next axiom. This is material for current research.

Emotion: I Can Evaluate the Results of Planning Different Actions According to Previous Experience

Part of the state machine evaluates the states found in "Virtual Machine Functionalism" section in terms previously obtained rewards or discouragements. The evaluation can create conflicts in "Volition: I Can Select What I Want and Can Act to Obtain It" section which are, sometimes resolved arbitrarily. This can give the organism a certain feeling of freedom. This too is a topic for current research.[5]

Questions That Need to Be Asked

In the spirit of a summary, it is now possible to return, from the perspective of the above guide, to some important

[5] Lee has a PhD thesis (London University) in preparation on this topic: *Aspects of affective action choice: computational modelling.*

questions that impact on the future of machine consciousness and address some potential scepticism.

What Licence Exists For Any Designed or Evolved, Non-Living Organism to Be Called 'Conscious'?

Clearly, no such license exists. There is also nobody who issues such licenses. It is up to the consciousness engineer to quote a benefit to understanding or application which justifies the use of the *machine consciousness* phrase. It is wrong to pretend that a machine said to be conscious is conscious like a living organism, not to mention a human. But respecting the distinction between living and 'artificial' conscious objects is instructive as it is possible to investigate what it is for the machine to be conscious of being a machine. By the same token, it is important to recognise the similarities between the phenomenal states in a machine and those we discover introspectively. A comparison can give us a quality measure for the success of the constructed machine.

Why Should a Conscious Artefact Have Advantages Over a Non-Conscious One, Where Behaviourally They May Be Indistinguishable?

It is not always evident that such advantages exist, but it needs to be stressed that while during a period of assessment a conscious and a non-conscious system can have identical behaviours, such behaviours may have been created in different ways, where a designer may claim that the conscious approach has advantages over a totally rule-controlled system without phenomenology. This was the aim of the parts of this paper relating to phenomenology. Also in robotics, there are opportunities to achieve greater autonomy, adaptation and learning stemming from the presence of phenomenal states. This has a practical edge over what can be achieved with classical rule-based cognitive systems where too many contingencies may have to be foreseen by a programmer.

Does Using the Language of Consciousness Have Engineering Advantages?

Here are two examples of when this is true. The first is to use the word *emotion* instead of something like "goal-centric self-evaluative capabilities that let the system self-manage its planning."[6] The other example is the use of the concept of a *phenomenal* state, which is evoked by the language of the study of consciousness as phenomenology as practiced at the turn of the last century: that is, with a first person representation at the heart of any mental process.

Are There Formal Ways of Discovering Consciousness In a System?

As mentioned above, this is not possible from the measurement of behaviour, as any conscious behaviour can be simulated purely by a sequence of instructions. However, in some of the author's work on phenomenal machine consciousness it was found useful to make the phenomenal states explicit (displayed on a computer screen). This allows qualitative judgements to be made on these states as compared to one's own introspection. There are examples of other approaches where quality measures of the density of interconnections may be introduced to show that a threshold needs to be exceeded to retain state structures significantly complex for consciousness.[7]

Can Machine Consciousness Be Studied Without Considering Phenomenal states?

Increasingly, the answer here is seen to be negative. Gamez [19], for example, defines consciousness as "the presence of a phenomenal world". There is a growing belief that those who use entirely functional methods rooted in AI must at least explain in what sense their models can be said to contain a phenomenal world, otherwise their work would not be considered as contributing to the aims of machine consciousness. Franklin et al. [20] show how such an argument may be conducted in the case of Global Workspace Theory through the addition of a "stable, coherent perceptual field".

Are There Some Computational Theories That Specifically Address Machine Consciousness?

If phenomenal states are to be taken seriously, fine grain computational techniques (i.e., neural networks) are necessary. Within this there is a choice of working with digital systems as shown in this paper, conventional neural networks (as in Haikonen) or spiking neurons (as in Gamez [19] and many others). It has also been argued that the computational concept of virtuality helps both with doing computational work on neurological models without recourse to clinical facilities while using clinical data, as well as providing freedom in the choice of physical substrates used in the modelling process.

[6] The author is grateful to Ricardo Sanz of Madrid Universtiy for this example.

[7] Tononi [17] has developed a set of necessary 'information integration' measures that are meant to be necessary for consciousness. Whether this measure indicates the presence of consciousness or not is being debated. See also the paper by Seth [18].

Do Specific Neuro-Scientific Theories Help in the Design of Conscious Systems?

The literature in neuroscience and consciousness is vast and the consciousness engineer should be aware of it even if it does not immediately dictate some salient design principles. Concepts such as the presence of brain areas that contribute to subjective feelings and the way they are supported by areas that do not, provide useful principles for the development of explanatory models.

Does Machine Consciousness Support New Philosophical Positions?

Again virtualism is important in showing that difficulties in both physicalism (the belief that mind and brain are the same thing) and dualism (the belief that mind and brain are either not connected, or only weakly so) may be overcome through reference to a well-understood flexible relationship between structure and function as found in computation.

Brief Conclusion

It has been argued in this paper that approaching consciousness as a design and modelling procedure using computational methods has the makings of both an explanatory methodology and the potential for the design of new systems. In the former case, models have included phenomenal internal states that stand in relation to the structure of the system that can be taken as an explanation of how phenomena might occur in the brain. Also this shows how a conscious mind as a virtual object may rely on a bounded infinity of physical structures. On the applied side, it has been seen that the design of robots may benefit in ways not afforded by more classical AI methodologies. But none of his means that all the work in this area has been done. On the contrary, the 'guide' character of this paper has only identified thin starting threads that those interested in designing conscious systems might care to follow or use as a foil to develop the paradigm in new directions.

References

1. Chalmers D. The conscious mind: in search of a fundamental theory. Oxford: Oxford University Press; 1996.
2. Penrose R. The emperor's new mind. Oxford: Oxford University Press; 1989.
3. Baars B. A cognitive theory of consciousness. New York: Cambridge University Press; 1988.
4. Shanahan M. Cognition, action selection and inner rehearsal. In Proceedings IJCAI workshop on modelling natural action selection; 2005. p. 92–99.
5. Dehaene S, Naccache L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition. 2001;79:1–37.
6. Haikonen P. The cognitive approach to conscious machines. Exeter, UK: Imprint Academic; 2003.
7. Haikonen P. Robot Brains: circuits and systems for conscious machines. Chichester: Wiley; 2007.
8. Block N. 'What is functionalism?'. The encyclopedia of philosophy supplement. New York: Macmillan; 1996
9. Sloman A, Chrisley R. Virtual machines and consciousness. J Consciousness Stud. 2003;10(4–5):133–72.
10. Dennett D. Consciousness explained. Boston: Little, Brown; 1991.
11. Holland O, Knight R, Newcombe R. A robot-based approach to machine consciousness. In: Chella A, Manzotti R, editors. Artificial consciousness. Exeter, UK: Imprint Academic; 2007. p. 156–73.
12. Hesslow G. Conscious thought as simulation of behaviour and perception. Trends Cognit Sci. 2002;6(2002):242–7.
13. Chella A, Frixione M, Gaglio S. Planning by imagination in CiceRobot, a robot for museum tours. In Proceedings of the AISB 2005 symposium on next generation approaches to machine consciousness: imagination, development, intersubjectivity, and embodiment; 2005. p. 40–49.
14. Husserl E. Logical investigations. London: Routledge; 1973 (English translation by JN Findlay, Original in German; 1901).
15. Aleksander I, Morton HB. An introduction to neural computing. London: Chapman and Hall; 1990.
16. Aleksander I. The world in my mind, my mind in the world: key mechanisms of consciousness in humans, animals and machines. Exeter, UK: Imprint Academic; 2005.
17. Tononi G. An information integration theory of consciousness. BMC Neurosci. 2004;5:42.
18. Seth A. Explanatory correlates of consciousness: theoretical and computational challenges. Cogn Comput. (this issue). doi: 10.1007/s12559-009-9007-x.
19. Gamez D. The development and analysis of conscious machines. University of Essex, PhD thesis in computing; 2008.
20. Franklin S, Baars BJ, Ramamurthy U. A phenomenally conscious robot? APA Newslett 2008;2(2):2–4.