

Owen Holland

## *Editorial Introduction*

In May 2001, the Swartz Foundation sponsored a workshop called ‘Can a machine be conscious?’ at the Banbury Center in Long Island ([http://www.swartzneuro.org/banbury\\_2001.cfm](http://www.swartzneuro.org/banbury_2001.cfm)). Around twenty psychologists, computer scientists, philosophers, physicists, neuroscientists, engineers, and industrialists spent three days in a mixture of short presentations and long and lively discussions. At the end, Christof Koch, the chair, asked for a show of hands to indicate who would now answer ‘Yes’ to the question forming the workshop theme. To everyone’s astonishment, all hands but one were raised. We had not asked the question at the beginning, and so we did not know if any minds had changed during the workshop, but I think we all realized the significance of this near-unanimous vote: the idea of machine consciousness had progressed from being an interesting philosophical diversion to a real possibility.

Later that year, the editors of the *Journal of Consciousness Studies* agreed that the topic would be suitable for a special issue of the journal, and submissions were invited from some of the Banbury workshop participants, and from others interested in the subject. I am grateful to all of the contributors for their co-operation and collaboration in bringing this collection together, and to the referees for the care with which they undertook their task. Special thanks go to Joseph Goguen, editor-in-chief of the *JCS*, and to managing editor Anthony Freeman for his patience and assistance throughout the project.

**Igor Aleksander** has spent several years engineering artificial neural systems to investigate and demonstrate various aspects of visual consciousness, particularly those involving imagination and imagery. One consequence is that he has probably spent more time than anyone else discussing and defending the notion that a machine might possess at least some of the attributes of consciousness. In their contribution to this collection, he and **Barry Dunmall** do not present a new neural model, but instead propose an axiomatic framework within which the structural and functional components of conscious systems, natural or artificial, can be identified and tested. They note: ‘We deem this to be useful if there is ever to be clarity in answering questions about whether this or the other organism is or

Correspondence: Owen Holland, Department of Computer Science, University of Essex, Wivenhoe Park, CO4 3SQ, U.K. *Email:* [owen@essex.ac.uk](mailto:owen@essex.ac.uk)

is not conscious.’ They emphasize that their approach ‘is meant to be open-ended’, so that others can contribute ‘further axiomatic clarifications’. Their current systems, embedded in robots, satisfy only three of their five axioms, and are therefore non-conscious, but within their formalism they are now able to ask: ‘[G]iven the development or evolution of the remaining two axiomatic mechanisms, what arguments could be used to deny them consciousness?’

Almost all the engineers and computer scientists involved in machine consciousness take a more or less conventional computational or neurally inspired approach, concentrating on the functions associated with cognitive processing. **Susan Blackmore**’s paper should give them pause: she suggests that our distinctively human consciousness centred on an experiencing self is an illusion created by the memes which have shaped our minds, and that the primary requirement for a machine ‘to think it was conscious’ is the ability to host memes — that is, to possess a capacity for imitation. This is usually low down on the list of cognitive abilities considered for implementation in artifacts (though there are signs that this is changing — see Nehaniv and Dautenhahn, 2002). Put bluntly, Blackmore appears to be saying not just that we might have missed something, but that we might have missed almost everything that matters. In support of her case she advances a wide variety of arguments, ranging from robotic experiments to evidence from meditation; she also identifies some key unanswered questions, asking in particular ‘whether artificial meme machines can ever transcend the illusion of self consciousness’.

In his paper on his new project, CyberChild, **Rodney Cotterill** brings together a number of approaches to the problem of machine consciousness. His chosen method is the computer simulation of the brain, body, and environment of a very young infant; the architecture of the child’s brain is a close neural model of what he has identified as the relevant parts of the mammalian nervous system; and the strategy is developmental and interactive, in that the child must signal its needs to the experimenter — for example, by crying appropriately — and the experimenter must respond. Cotterill is very open minded: although he has a well developed theory of consciousness, he makes it clear that his current project is broadly investigative, ‘searching for the neural correlates of consciousness through computer simulation’ rather than explicitly testing any single narrow hypothesis. CyberChild possesses not only a simulated brain and body, but also a simulated metabolism; learning to deal with the contingencies presented by its environment is a matter of life and death, and by implication the approach emphasizes the functional links of emergent consciousness to the well-being of the organism. Although the simulation is necessarily much less complex than the reality it is intended to mimic — as Cotterill puts it, ‘In CyberChild, one sees the nervous system pared down to its essentials’ — the resultant simplicity offers the advantage that ‘If evidence of conscious behaviour does emerge . . . one could be reasonably optimistic that its neural correlates will be detectable.’ The work seen in close focus is rooted in biology, but Cotterill expresses the hope that success will constitute ‘a step toward realizing the long-cherished dream of creating *Homo siliciens*: consciousness in a computer’.

In many ways, **Stan Franklin**'s work on 'conscious software' offers a real challenge to functionalists. If consciousness is what consciousness does, then his systems may well exceed the requirements, in that they not only mimic successfully the outcomes of conscious processes in some humans (naval despatchers) but they do it in the way that the conscious human brain appears to do it, since their functional components are explicitly modelled on the elements of Baars' global workspace theory (Baars, 1988). Franklin is happy to describe his system IDA as having functional consciousness (in a sense which he defines carefully); in this paper, he goes on to speculate about what more might be said about the system, and what more might be added to extend and perhaps deepen any consciousness that exists in the current version. He is in a privileged position: if IDA is judged not to be conscious because 'she' lacks some functional or structural component X (such as a self), then he and his team may be able in due course to add X. What then? As with Aleksander and Dunmall, who offer their axioms up for extension and modification if anyone feels they are deficient, the onus is on the critic to characterize what is lacking — and if the critic obliges in precise enough terms, it may be possible to correct the deficiency. From a technical point of view, IDA occupies a curious position: she is disembodied, but performs a real-world task in real time. She is also extremely complex. Embodiment and complexity are two key themes raised in other papers in this collection.

In his characteristically vigorous piece, **Stevan Harnad** tackles the issue of how we might know that a machine — an artifact — was conscious, and argues for 'behaviour-based Turing testing' as being necessarily at the root of all our attributions of consciousness, even to other humans. When Johnson remarked of an evening spent in a tavern with company that, 'We had good talk,' Boswell commented, 'Yes sir, you tossed and gored several persons.' Harnad does the same with several ideas, revealing the behavioural nature of the evidence underlying judgments of both the presence and correlates of consciousness. However, advocates of machine consciousness who find their optimism rising as they read the essay may be sobered by the last paragraph, where Harnad reminds us that 'our forward- and reverse-engineering (of Turing indistinguishable robots) can only explain how it is that we can do, not how it is that we can feel.'

Like Cotterill, **Owen Holland** and **Rod Goodman** do not start with consciousness, but hope to end up with it at some future time. And like Blackmore they emphasize a single mechanism — internal modelling — as the possible underpinning of consciousness. (Internal modelling is not imitation, but the two notions are close enough to give food for thought.) Their approach is rooted in robotics; their claim is that a robot able to deal intelligently with the complexities of the real world will have to engage in planning, and that this requirement will inevitably demand the creation of an internal model not just of the world, but of many aspects of the embodied agent itself. They speculate that such an internal agent-model may give rise to some consciousness-like phenomena. Their strategy, like Cotterill's, is developmental, but rather than allowing an entity to modify and extend its own capabilities, they propose to re-engineer the robot themselves, adding and changing whatever is necessary to deal with the

progressively more difficult environmental contingencies to which they intend to expose it. Like Aleksander and Dunmall, their starting point is a robot that they claim is definitely not conscious; from there, as they remark, ‘The only way is up.’

At the end of his paper, after presenting and defending a version of mysterianism, **Jesse Prinz** offers some advice to engineers: ‘[E]ngineers should continue trying to model what they can, and they should stop trying to model what they can’t. . . . [They] shouldn’t fool themselves into thinking they can definitely create conscious machines.’ If correct, this is surely good advice; no engineer wants to waste prime-time decades on an impossible quest (that’s the territory of the physicist.) The roots of his mysterianism echo the concerns of the single dissenter at the Banbury meeting: the impossibility of being sure that the biological substrate does not contain, at some low level, properties essential for consciousness. After developing his positions on two of the usual objections to machine consciousness, Prinz then goes on to present a theory of consciousness which he uses ‘to show that progress in the science of consciousness may offer little help to those who want to engineer consciousness’. Like Harnad, he is concerned that behavioural evidence is at the root of attributions of consciousness, but unlike Harnad, he restricts his concern to dealing with ‘inorganic brains’, and this is why he is sceptical. Engineers might find a hint of optimism in his statement of what he is sceptical about: ‘The problem isn’t that it would be impossible to create a conscious computer. The problem is that we cannot know whether it is possible.’

Rather than working from an understanding of consciousness to the construction of a machine to support it, **Aaron Sloman** and **Ron Chrisley** offer a different approach. Arguing that our existing concepts of mind are ‘pre-theoretical’, they consider progressively more complex information processing architectures based on virtual machines, and use the possibilities offered by various distinct types of architectures both to illuminate some of the different meanings of consciousness, and to suggest more useful and accurate concepts. In the process, they deal with many of the issues which face the designer of any intelligent agent, and provide what is sure to be a useful taxonomy of mechanisms and possible systems. An example of the potential power of their approach can be seen in their discussion of qualia (a topic usually avoided by engineers) within what is essentially an engineering framework. Their paper does not propose the construction of any particular machine — the notions used are abstractions, rather than concrete suggestions for implementation. Abstraction usually implies a degree of simplification, but one of the striking features about their work is the complexity of the architectures they are able to generate.

**Luc Steels** offers a view of a key feature of consciousness — the inner voice — from the perspective of research into the acquisition and use of language by artificial agents. He expresses his methodological stance as follows: ‘[W]hatever consciousness “really” is, some of the behaviours often associated with having consciousness can be unravelled, and their information processing foundations understood.’ He describes how a community of agents can readily arrive at a

shared lexicon by engaging in a robotic ‘language game’, but how ‘the emergence of grammar has turned out to be much more difficult’. This problem was solved by the introduction of a particular information processing strategy — a re-entrant mapping, when output from the language production system is fed back into the language interpretation system. He describes how this enabled the acquisition of a form of grammar by the members of a suitably programmed agent community. However, this re-entrant system (which would not be out of place in one of Sloman and Chrisley’s architectures) also has the potential for enabling a range of entirely new processes, including an inner voice that could provide the foundation for the construction and testing of a self-model. Although such a self-model ‘is not to be identified with consciousness’, he argues that it is part of the conscious experience, and so the development of language in the way he describes ‘may have played a crucial role in the origins of consciousness’. Although he is optimistic about the future progress of robotics in capturing some of the information processing aspects of consciousness, he leaves the question of producing first person experience open.

As one might expect, our last contributor, **William Irwin Thompson**, deals with machine consciousness from a rather different point of view, by examining the cultural causes and implications of its pursuit in the West at this particular time. He paints a dark picture. He wrote in 1992, ‘There seems little chance of getting out of this century with the same human nature with which we entered it’ (Thompson 1992); his concern now is that ‘in order to grant consciousness to machines, the engineers first labour to subtract it from humans’, and ‘the humanistic philosopher of mind . . . finds himself replaced by the robotics scientist’. Interestingly, some of Thompson’s observations concerning the mistaken path of ‘the mechanists’ resonate with some of the papers in this collection. He emphasizes that ‘slowness is fundamental to the nature of consciousness’, and that the ‘slow, sloppy’ nature of nervous tissue is responsible for this; both Harnad and Prinz also consider the question of whether the biological substrate may hold some property essential to consciousness. In a passage speculating on the evolution of consciousness, he examines the problems of an organism possessing ‘multiple channels of sensory registration’; it is difficult to resist the temptation to locate the creature within Sloman and Chrisley’s taxonomy of information processing architectures. However, Thompson is deeply suspicious not only of the methods but also of the motivations of ‘the mechanists’: he thinks they (and we?), caught up in hypercapitalism, are simply ‘hawking their wares’. But he also holds out the hope that monolithic corporate capitalism and American techno-idolatry may not in the end prevail — that they may only delay ‘our emergence to an enlightened planetary culture’ for a century or two, or even be outcompeted by a new form of capitalism ‘wed to information technologies and complex dynamical systems’.

It is now fourteen years since the publication of Leonard Angel’s book *How to Build a Conscious Machine* (Angel, 1989) — perhaps the first serious consideration of making a practical assault on the problem of machine consciousness. The book has worn well — some might take this as an indication that progress

has been slow. However, it is interesting to note some of the differences between it and Pentti Haikonen's recent book *The Cognitive Approach to Conscious Machines* (Haikonen, 2003). Angel, a university-based philosopher interested in artificial intelligence, wrote from a philosophically-influenced perspective; his main concern was to shed light on 'the traditional mind/body problem'. Haikonen is an engineer working for a major technology company; his text contains system block diagrams, signal flow diagrams, and visual subsystem diagrams, and his preface mentions that he is already working on 'neuron group microchip development for the eventual implementation' of his machines. This gradual shift from the armchair to the laboratory and the workshop can also be seen, I believe, in the present collection. We cannot yet know how fast and how far the enterprise will progress, and how much light it will be able to shed on the nature of consciousness itself, but it seems beyond doubt that machine consciousness can now take its place as a valid subject area within the broad sweep of consciousness studies.

### References

- Angel, L. (1989), *How to Build a Conscious Machine* (Boulder, CO: Westview Press).  
Baars, B.J. (1988), *A Cognitive Theory of Consciousness* (Cambridge: Cambridge University Press).  
Haikonen, P.O. (2003), *The Cognitive Approach to Conscious Machines* (Exeter: Imprint Academic).  
Nehaniv, C. and Dautenhahn, K. (2002), *Imitation in Animals and Artifacts* (Cambridge, MA: MIT Press).  
Thompson, W.I. (1992), *The American Replacement of Nature: The Everyday Acts and Outrageous Evolution of Economic Life* (New York: Doubleday).