

Robert Clowes, Steve Torrance  
& Ron Chrisley

## *Machine Consciousness*

### *Embodiment and Imagination*

Readers of this Journal are used to considering questions to do with consciousness and subjectivity in humans and other natural creatures. However it is instructive to devote some attention to the realm of artificial subjectivity — as a bare possibility, even if not as a likelihood in the near future. Following the first special issue of this journal on the topic of machine consciousness (Holland, 2003) there has been a lot of further interest in the topic: a topic which covers, not just the development of consciousness and subjectivity in machines, but also the use of machine models — computer software, robots, and so on — to help shed light on consciousness as a wider phenomenon. The latter pursuit seems a valuable extension to the use of computer models more broadly in cognitive science. The former goal — hubristic and ethically dubious as some might see it — raises intriguing theoretical and philosophical puzzles. The two research strands, although logically separable, are harder to disentangle in practice, as will be seen in the contributions below.

We will not rehearse here the early history of the field of machine consciousness (MC) as it is ably articulated both by Holland in his introduction to the 2003 *JCS* volume, and in a paper in this issue co-authored by one of its founding practitioners (**Aleksander & Morton**). Suffice it to say that most of the papers included here were first presented at one of two recent two-day workshops on Machine Consciousness that were held in Hatfield (Chrisley *et al.*, 2005), and Bristol (Clowes *et al.*, 2006). With one exception, all the principal authors here spoke at one or both of these workshops, both of which were sessions in annual conventions of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB).

Two key themes that emerged from these workshops were *embodiment* and *imagination*. These play a central role in many of the papers below. We see these two areas of inquiry as unifying much current work in the field of machine consciousness. We will briefly consider how far these two themes get us towards building machines that are actually conscious; and how pursuing such a goal — and the less ambitious goal of using machines to model consciousness (MMC)<sup>1</sup>— might relate to consciousness studies more generally.

### Embodiment

‘Embodiment’ is a term that can be understood in many ways. In his contribution, **Holland** calls for a ‘strongly embodied approach to machine consciousness’ and his discussion of the anthropomorphic robot CRONOS built for this purpose (and illustrated on our cover) goes some way towards indicating what such an approach involves. First, it requires of a potentially conscious robot that its physical instantiation — especially its means of movement — strongly resembles the means and modalities of movement available to human beings. However, the second part of Holland’s account emphasizes (in line with Metzinger, 2004) a requirement for *self-modelling* as central for embodied, motor-oriented, subjectivity, machine or otherwise<sup>2</sup>. So, alongside building CRONOS, Holland and his colleagues are also building a virtual model, SIMNOS, through which CRONOS’s awareness is mediated.

**Holland’s** approach to embodied consciousness seems to be strongly compatible with **Kiverstein’s** Dynamic Sensorimotor (DSM) account of consciousness which argues, following Hurley (1998) and O’Regan & Noë (2001), that consciousness arises from the exercise of the mastery of sensorimotor regularities. **Kiverstein’s** complex and subtle argument aims to demonstrate that the DSM account can show how an artificial agent that exercises the appropriate sensorimotor knowledge has a subjective point of view, and hence a consciousness of itself as the owner of experiences. Such an approach implies that the fine-grained character of motor control has a key significance for consciousness (although see Clark, 2001, for some qualifications).

A second sense of embodied self may be related to **Aleksander and Morton’s** first axiom of *presence* (previously referred to as ‘being in

- 
- [1] Where the context does not make clear which of the two strands we are referring to at any time, we will in this introduction use the term ‘Machine Modelling of Consciousness’ (MMC) to refer specifically to the creation of models of consciousness (following Aleksander, 2007), and MC to refer to the project of creating consciousness *in* machines, (which will of course invariably include MMC).
- [2] Which leads us to our second theme, imagination, to be discussed below.

an out-there-world'), which is characterized in their article as 'perception plus a sense of *being*'. In fact, most authors in this volume indicate that building in some sense of embodied presence would be central to the task of the construction of a conscious machine.

**Ziemke** proposes, however, that despite an apparent convergence towards an embodied approach in the MC community, there remains much disagreement about what is involved in embodiment. He suggests that the field is split into two main camps. One, exemplified in the papers by **Holland, Kiverstein** (as seen above), and also **Hesslow & Jirehned, Ikegami** and **Haikonen**, emphasizes complex dynamics of sensorimotor engagement as being central to the relevant notion of embodiment. This notion of embodiment goes well beyond traditional 'in-the-head' conceptions of subjectivity, but is still highly abstract: it does not commit to anything that implies a particular kind of organic constitution.

A contrasting position, which **Ziemke** develops in some detail, understands embodiment more richly, as being more biologically oriented, and further, as being bound up with the notion of a 'lived body', i.e. a body that is an organised autopoietic unity along the lines specified by Varela and others, and that has been developed subsequently within the enactive approach (Maturana & Varela, 1987; Varela *et al.*, 1991). **Torrance** also outlines an enactive conception of embodiment as integral to what he calls 'thick phenomenality'. The autopoietic conception of lived embodiment is proposed by him as one promising way to develop the notion of thick phenomenality, but his arguments in favour of the latter also depend upon weaknesses perceived in the alternative, 'thin', conception that he identifies. Both **Torrance** and **Ziemke** see progress on the MC project as much harder, but not impossible, for all that, on the richer conceptions of embodiment and phenomenality that they present.

**Stuart's** paper proposes another way of understanding embodiment. Her account draws on a number of inspirational sources, which she sees as all involving a number of common elements: 'the notions of engaged embodiment, goal-directed animation, perception, and imagination'. **Stuart's** account also implies a rich sense of embodiment, that is perhaps midway between the relatively abstract notions based on dynamic sensorimotor knowledge, and the autopoietic conceptions discussed by **Ziemke** and **Torrance**. Like them she sees her account as making it harder, but not impossible to realize consciousness as an engineering project.

Despite the importance given to embodiment in many of these papers, not all authors agree that it is central. Another line of thought is that, as well as (or maybe rather than) the right kind of embodiment,

a conscious machine would require an ‘inner world’ and this seems to be often taken to be synonymous with having imagination.

### Imagination and the Inner World

Building robots or other artefacts which can be said to have *an inner world* or a capacity for *imagination* in some sense, has been a particular focus of MC research which has set it apart from anything found in more standard artificial intelligence or robotics research. **Stuart**, as we have seen, makes embodiment central to consciousness, but she also argues that a ‘synthetic’ imagination — conceived of in a way that leans heavily on the work of Immanuel Kant — would crucially be required for a genuinely conscious machine. At least two important senses of imagination are at play here: the notion of cognitive imagination, which occupies a central role in Kant’s account of the synthesis of experience; and that of tactile-kinaesthetic imagination which, **Stuart** argues, is central to how a creature becomes aware of itself, through movement, as a embodied creature in a world.

While there is much agreement on the centrality of imagination to the MC project, there remain questions about what it consists in and how it relates to consciousness more generally. According to **Aleksander & Morton**’s axiomatic approach, if presence (first axiom) provides the basic form of experience then imagination (second axiom) is ‘a more or less degraded version of that experience’ that allows a sort of derivative or virtual presence. For **Haikonen**, however, ‘our senses produce only a limited amount of information, which has to be augmented by imagination’. The role of imagination in **Haikonen**’s work is also different in that global availability of anticipated or actual information is emphasized, a feature that will remind many of the Global Workspace Hypothesis (Baars, 1988). Also, like **Holland**, **Haikonen** stresses the importance, for consciousness, of the *transparency* of representation (a subject’s access to the content, but not the vehicle, of information). Readers may wish to consult his previous and forthcoming books on machine consciousness in order to get a better idea of how his architecture is intended to accommodate transparency more fully than the approaches of others.

One of the most detailed attempts to argue that a conscious robot must be considered to have an inner world is found in **Hesslow & Jirenhd**. Their approach rests on **Hesslow**’s *simulation hypothesis* (Hesslow, 1994) which argues that what we mean by saying that an agent has an inner world is that it is able, to some sufficient degree, to simulate its interaction with the external world. **Hesslow & Jirenhd**

argue that their simple, idealized robot *K*, in virtue of being able to use its simulation capabilities to imagine interactions with the real world, can be argued to have an inner life, and hence a rudimentary consciousness. **Clowes** questions the sufficiency of the minimal approach to imagination by focusing on how imagined speech might come to play a role in inner life. While following **Hesslow's** basic schema about what imagination — if not consciousness — is, **Clowes** argues for an internalisation approach to consciousness which emphasizes the functional reorganisation which takes place alongside the establishment of a system of inner rehearsal.

Imagination is also a central issue in **Ikegami's** contribution, which concerns a minimal agent whose dynamics propel it to switch periodically between a more environment-oriented sampling activity to one driven by its own evolving internal dynamics. The model underlying these experiments involves neural units which have a complex time signature. An agent activity pattern is generated, which shifts between seemingly random environmental exploration and exploration of its internal dynamics. **Ikegami's** model suggests a possible convergence of the themes of embodiment and imagination here, as his minimal robots shift between inner focused and outer (perceptual) modes of organisation as a natural function of the neural implementation.

**Chrisley & Parthemore** offer another interesting perspective on understanding the role of imagination in the MC project. They believe their approach — characterized as 'synthetic phenomenology' — might allow us a way to use machines to specify some of the fine-grained structure of consciousness while remaining uncommitted as to the possibility of achieving genuinely conscious machines. Their goal is to develop a robotic platform to attempt to specify the detailed structure of mental states in a way that it is difficult or impossible to state in words. Their model (implemented on a Sony Aibo robot) enables phenomena such as change-detection in a scene and forms of foveation, change-blindness, etc. to be reproduced. This programme of machine phenomenology<sup>3</sup> may prove to be complementary to other MC work. Their work can also be seen as uniting the two themes of embodiment and imagination: their approach exploits the embodiment and situatedness of a robot to specify experiential states; and they assume a sensory-motor approach that takes consciousness to consist in a capacity to anticipate, or imagine, the sensory input that would be received were one to move this way or that.

---

[3] Not to be confused with machine phenomenality.

### Minimal Subjectivities and Success-conditions for MC

**Hesslow's** simulation hypothesis, **Kiverstein's** DSM approach and even **Ikegami's** embodied chaotic itinerancy all imply different minimality criteria for subjectivity. **Kiverstein's** minimal subjectivity just needs to practice a mastery of its sensorimotor skills in order to be minimally conscious, while **Hesslow's** robot K in addition simulates its sensorimotor interactions. For **Ikegami**, spontaneous fluctuations between active perception and inner directedness are necessary, while for **Holland** being a comprehensive self-simulator is required. The detail of these approaches shows that there is indeed a fair amount of common ground here. Nevertheless questions can also be raised as to whether these approaches can really be reconciled.

Perhaps the most ambitious theoretical attempt at such integration is the axiomatic approach of **Aleksander & Morton**, first put forward in Aleksander and Dunmall, 2003. In the version presented here<sup>4</sup> there are clearer connections to a phenomenological tradition, as well as further elucidation of a proposed core architecture which would integrate systems to deal with different core functions. **Haikonon** suggests, however, that there is a need to develop more specifically human-like minds — in particular minds that have reflective imaginations, inner speech and complex forms of self-awareness. **Haikonon's** 'system' approach also makes inter-modal communication and reportability central to the requirements of a truly conscious machine.

There remains, of course, some scepticism on the likely outcome of the MC project. **Bringsjord** expresses a deep-seated scepticism, based on his view that (notwithstanding **Aleksander's** proposed axioms) no clear, formal criteria for consciousness have been offered, and on his suspicion that no such criteria could be provided. **Bringsjord's** reserve about MC does not stem from a blanket reserve about AI projects *per se*. He is quite happy to accept such projects when they permit implementation using the formal methods that he champions, and indeed he cites AI work by himself that uses such methods.

**Torrance** has different grounds for scepticism, although his scepticism is perhaps less marked than **Bringsjord's**. On the former's view, much MC research fails to acknowledge the centrality of phenomenality to consciousness, or to appreciate the different possible conceptions of phenomenality that are available — for example 'thin' versus 'thick'. The 'thin' conception sees consciousness as a kind of super-layer upon physical or functional aspects of an agent. Much scepticism about MC, and indeed physicalist approaches to

[4] The axioms are now called *presence, imagination, attention, volition* and *emotions*.

consciousness in general, is based on variants of the ‘thin’ conception. But some supporters of MC have implicitly relied on a ‘thin’ conception as well. This allows them to claim conditions of success for MC which are rather too easy to fulfill (e.g. by allowing relatively simple functional models to be declared as being potential instantiations of consciousness). The alternative, ‘thick’, conception, by incorporating a rich conception of embodiment (see above), sets much more challenging success-conditions.

As we have seen, ‘Machine Consciousness’ comprises at least two different themes of study — the use of computational, robotic and other artificial means to model consciousness in order to understand it better (MMC); and the attempt to create consciousness artificially. (The latter might be called ‘strong MC’ following Searle’s [1980] distinction between weak and strong AI [cf. **Torrance**].) Several of the papers focus on work in progress by the authors on the MMC enterprise (**Aleksander & Morton, Chrisley & Parthemore, Clowes, Haikonen, Hesslow & Jirenhed, Holland, Ikegami**). Some at least of those reporting current MMC work in progress consider themselves as also addressing the strong MC question. For example **Hesslow & Jirenhed** see no as yet undiscovered component as needed in order to realize full-blooded MC, it being just a matter of adding increased complexity to existing system design methods.

However most of those who focus specifically on the strong MC project in this issue adopt a more theoretically-driven perspective. Of these, **Kiverstein**’s arguments imply a sharp optimism about realizing strong MC — in principle, and perhaps even in a reasonable timeframe. **Stuart**’s position also seems to be fundamentally optimistic, although she offers some fairly tough criteria for a genuinely conscious machine. **Ziemke** and **Torrance** are rather more guarded; while **Bringjord**’s paper offers an uncompromising challenge to anyone seriously embarking on a strong MC programme.

It is worth reflecting on how MC (and MMC) research issues interconnect with mainstream consciousness debate. Many writers on consciousness who are not part of the MC community as such have expressed optimism about the strong MC programme at least in principle (Dennett, Chalmers, Baars, Metzinger and O’Regan might be mentioned). Also there is the question of the coarseness or fineness of grain in the theory of the physical conditions that will provide a sufficient explanation for how consciousness arises — clearly this impacts on how close or distant the technical realizability of artificial consciousness is seen to be. Discussions over MC impact on many other mainstream issues concerning consciousness: how consciousness is to

be defined or characterized; what different kinds of consciousness there are (core, phenomenal, access, functional, ...); how these different types relate to one another, and so on.

Finally, it is worth remembering that questions concerning artificial consciousness and artificial mentality more generally have important social and ethical ramifications, as more and more robots and other kinds of artificial humanlike agents are produced. Our predominating social and moral attitudes may be transformed as such technologies proliferate — indeed there may be tectonic shifts in prevailing notions of what ‘society’ is, and who ‘we’ are. So, in decades to come, developing work on machine consciousness may come strongly to affect how consciousness is seen, by both lay people and experts alike.

### *Acknowledgements*

The editors thank Margarita Stapleton, whose organisational prowess was essential in running the Bristol AISB workshop; and Joel Parthemore for invaluable help at key stages in preparing of this special issue; also the conference chairs for the AISB-05 and AISB-06 conventions, and all who acted as anonymous referees, both for the workshops and for volume. Finally we thank Anthony Freeman for his patience and generous help at every stage in the production of this issue.

### **References**

- Aleksander, I., & Dunmall, B. (2003), ‘Axioms and tests for the presence of minimal consciousness in agents’, *Journal of Consciousness Studies*, **10** (4), pp. 7–18.
- Aleksander, I. (2007), ‘Machine consciousness’, in *The Blackwell Companion to Consciousness*, ed. Velmans and Schneider (Oxford: Blackwell), pp. 87–98.
- Baars, B. (1988), *A Cognitive Theory of Consciousness* (Cambridge: CUP).
- Clark, A. (2001). ‘Visual experience and motor action: Are the bonds too tight?’ *Philosophical Review*, **110** (4), pp. 495–519.
- Chrisley, R., Clowes, R. and Torrance, S (ed. 2005), *Proceedings of the Symposium on Next-Generation Approaches to Machine Consciousness*, AISB-05, (Hatfield: University of Hertfordshire).
- Clowes, R., Chrisley, R. and Torrance, S. (ed. 2006), *Proceedings of 2006 Symposium on Integrative Approaches to Machine Consciousness*. In T.Kovacs and J.Marshall (eds). *AISB’06: Adaptation in Artificial and Biological Systems, Vol 1*. (Bristol: University of Bristol), pp. 107–73.
- Hesslow, G. (1994), ‘Will neuroscience explain consciousness?’ *Journal of Theoretical Biology*, **171**(1), pp. 29–39.
- Holland, O. (ed. 2003), ‘Editorial introduction’, *Journal of Consciousness Studies*, **12** (4–5), Special issue on Machine Consciousness, pp. 1–6 .
- Hurley, S. (1998), *Consciousness in Action* (Cambridge, MA: Harvard UP).
- Maturana, H.R. & Varela, F.J. (1980), *Autopoiesis and Cognition* (Dordrecht: Reidel).
- Metzinger, T. (2004), *Being No One* (Cambridge, MA: MIT Press).
- O’Regan, J.K. & Noë, A. (2001), ‘A sensorimotor account of vision and visual consciousness’, *Behavioral and Brain Sciences*, **24** (5), pp. 939–73.
- Searle, J.R. (1980). ‘Minds, brains, and programs.’ *Behavioral and Brain Sciences*, **3**(3), pp. 417–57.
- Varela, F., Thompson, E. & Rosch, E. (1991), *The Embodied Mind*. (Cambridge, MA: MIT Press).