References

Blakemore, C. (2005), 'Harveian oration: In celebration of cerebration', *The Lancet*, **366**, pp. 2035–57.

Chalmers, D.J. (1996), *The Conscious Mind* (New York: Oxford University Press).

Eccles, J.C. (1994), *How the Self Controls Its Brain* (Berlin:Springer-Verlag).

James, W. (1890), *The Principles of Psychology*, Chapter 5: The Automaton-Theory. This can be found at: http://psychclassics.asu.edu/James/Principles/prin5.htm

Plantinga, A. (2004), 'Evolution, epiphenomenalism, reductionism', *Philosophical and Phenomenological Research*, **68** (3), pp. 602–19.

Pockett, S. (2004), 'Does consciousness cause behaviour?', *Journal of Consciousness Studies*, **11** (2), pp. 23–40.

Popper, K.R. and Eccles, J.C. (1977), *The Self and Its Brain* (Berlin:Springer).

Robinson,W.S. (2007), 'Evolution and epiphenomenalism', *Journal of Consciousness Studies* ,**14** (11), pp. 27–42.

Swinburne, R. (1996), 'The soul', in *Philosophy of Mind*, ed. T. O'Connor and D.Robb (London:Routledge,2003). Originally Chapter 5 of *Is There a God?* (Oxford: Oxford University Press,1996).

Velmans, M. (1991), 'Is human information processing conscious?', *Behavioral and Brain Sciences*, **14**, pp. 651–726.

Igor Aleksander, Uziel Awret, Selmer Bringsjord, Ron Chrisley, Robert Clowes, Joel Parthemore, Susan Stuart, Steve Torrance and Tom Ziemke

Assessing Artificial Consciousness

A Collective Review Article

Background

While the recent special issue of *JCS* on machine consciousness (Volume 14, Issue 7) was in preparation, a collection of papers on the same topic, entitled *Artificial Consciousness* and edited by Antonio Chella and Riccardo Manzotti, was published.[1] The editors of the *JCS* special issue, Ron Chrisley, Robert Clowes and Steve Torrance, thought it would be a timely and productive move to have authors of papers in their collection review the papers in the Chella and Manzotti book, and include these reviews in the special issue of the journal. Eight of the *JCS* authors (plus Uziel Awret) volunteered to review one or more of the fifteen papers in *Artificial Consciousness*; these individual reviews were then collected together with a minimal amount of editing to produce a seamless chapter-by-chapter review of the entire book. Because the number and length of contributions to the *JCS* issue was greater than expected, the collective review of *Artificial Consciousness* had to be omitted, but here at last it is. Each paper's review is written by a single author, so any comments made may not reflect the opinions of all nine of the joint authors!

Correspondence:
Ron Chrisley, COGS/Dept of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, U.K. *Email: ronc@sussex.ac.uk*

[1] *Artificial Consciousness*, ed. A.Chella & R. Manzotti (Imprint Academic, 2007)

The Chapters Reviewed

It's entirely fitting that Vincenzo Tagliasco begins his survey of the history of artificial consciousness ('Artificial Consciousness: A Technological Discipline') with an anecdote about his own introduction to the field: consciousness is, after all, ultimately a personal affair. Descartes comes in for kind words; engineers are quite happy to get on with the metaphor of humans as machines, and leave the worries about dualism to the philosophers. That's Tagliasco's main point: researchers in the field are more engineers than theoreticians (though philosophers are welcome!). They want to build things and see what interesting properties they exhibit. Making a copy of human consciousness isn't on the table. Producing a robot that evolves into one with recognizably conscious behaviour, on the other hand, is. 'An artificial conscious being would', he writes, 'be a being which appears to be conscious because [it] acts and behaves as a conscious human being.'

What is 'artificial consciousness': artificial *consciousness* or *artificial* consciousness? Is it 'real' consciousness achieved by artificial means, or something that resembles consciousness in the way an arrangement of artificial flowers resembles the real thing? From a distance they're quite impressive; just don't examine them too closely! Agreeing on one set of terms or one approach in what is a young discipline is, Tagliasco believes, a distraction at best. The first step must be to make sure researchers aren't just talking past each other. The advantage of the engineering perspective is in putting theory into practice: 'technology', Tagliasco writes, 'overcomes ambiguity' — a point that philosophers might do well to remember!

John Taylor's paper, 'Through Machine Attention to Machine Consciousness', aims to make three contributions: a philosophical analysis of the architectural requirements for consciousness, a demonstration that a particular, independently-motivated, control-theoretic model of attention can meet these requirements, and a discussion of the specific issues that must be resolved in attempting to implement such a model in an artificial system such as a robot. Consciousness is taken to involve not only control of attention, but two other components, one concerning contentful, world-directed states, the other (following the phenomenological tradition) being the (contentless) pre-reflective self. The pre-reflective self is that which confers our ownership of our perceptions, that which 'gives us the sense of "being there", of "what it is like to be"'. It is this which, Taylor claims, allows our representations to mean anything to us. Taylor then presents the

CODAM (Corollary Discharge of Attention Movement) model, a control-theoretic architecture comprising a plant, goal modules, inverse models and forward models, supplemented with a sensory working memory buffer, and the corollary discharge buffer (WMcd), which is a prediction of the attended input given the attentional movement. It is shown how this model can explain various perceptual and attentional phenomena, such as those seen in the Posner movement benefit paradigm (Rushworth *et al.*, 1997) and the attentional blink (Vogel *et al.*, 1998) on the one hand, and pure conscious experience (Forman, 1999) on the other. This explanatory capacity is meant to establish a connection between WMcd and pre-reflective consciousness: modelling pure conscious experience establishes the contentless nature of the WMcd, while the anticipation of input used in explaining the Posner benefit and attentional blink is meant to confer the 'ownership' aspects of the pre-reflective self.

But this is left as a tantalizing suggestion, leaving the reader to wonder how the equation is supposed to be secured. Why would (correct) anticipation confer a sense of ownership on perceptions? The increase in reaction times associated with the attentional blink not withstanding, we still experience unexpected, unanticipated inputs as our own. On the other hand, the model seems too simple to be sufficient for consciousness — Taylor doesn't (here) make it clear why a complex thermostat couldn't implement CODAM. This leaves one wondering what else must be added to elevate CODAM from a model of some aspects of consciousness-related processing to actually being sufficient for experience, as (ambitious) machine consciousness requires. The final section enumerates some implementation issues that those attempting machine consciousness should consider (although the Searle-like argument for why CODAM will only succeed in producing consciousness if implemented in hardware rather than software is too brief to persuade anyone). Nevertheless, the model seems a very good place to start, and the questions it raises seem to be the right ones to ask. Of further interest is a handy table listing aspects of conscious experience and known aspects of the nervous system that seem to support such phenomena.

In 'What's Life Got To Do With It?', Tom Ziemke claims, and he is not wrong, that in our attempt to create embodied AI, autonomous agents, and artificial consciousness we have paid too little attention to theoretical biology and have not yet grasped the crucial role that the living body plays in the constitution of the self and of forms or aspects of consciousness. He claims that when 'we refer to both living organisms and robots as "autonomous agents", it is important to keep in

mind that their "autonomy" and "agency" are fundamentally different'. We should adopt a position of 'caveat spectator' and not take similarity of behaviour for similarity of underpinning. The underpinning, the biology, the internal constitution and regulation are crucial because 'the way an organism constructs itself also shapes the way it constructs its self'. Thus, he asks, what has life got to do with consciousness and the development of a sense of self?

In answer to this question he weaves together von Uexküll's organismic biology and the concept of autonomy, Maturana and Varela's theory of autopoiesis, and work in evolutionary and epigenetic robotics — all of which have at their heart the construction of knowledge 'in sensorimotor interaction with the environment with the goal of achieving some "fit" of "equilibrium" between internal behavioural/conceptual structures and experiences of the environment' — with Damasio's somatic theory and pre-conscious proto-self which is able to continuously map the physical state and structure of the organism in its dynamic engagement with its environment. It is this organisation, this self-construction and -preservation, that Ziemke now emphasizes, for it is this natural autopoiesis within an operationally open system of agent–environment interaction that is the source of more developed notions of core- and extended-consciousness.

The use of 'natural' to qualify 'autopoiesis' is not Ziemke's; his distinction is between auto- and allo-poietic systems, with man-made artefacts like robots and software agents conforming to the latter category for their 'components are produced by other processes that are independent of the organisation of the system'. And here is where a criticism of Ziemke's enterprise arises. There is an autonomy that an autopoietic system possesses and an allopoietic system lacks, and this autonomy is crucial for the development of consciousness in any system. But if there is no way for an allopoietic system to ever become an autopoietic one, then it would seem that the construction of truly conscious machines, other than biological machines, is beyond us, even if we have Maturana and Varela's assurance that autopoiesis is about organisation not the realising structure.

Aleksander is in characteristic form in the article 'Depictive Architectures for Synthetic Phenomenology', and not chary about embracing awkward concepts like 'phenomenology' and 'introspection'. Indeed the primary question asked here by Igor Aleksander and Helen Morton is whether phenomenology has any purchase in the computational domain. They argue towards a synthetic phenomenology resulting from the combination of two things: (i) the capacity for first-person ascription to the computational model or architecture, and

(ii) the model's ability to explain the action-usable representation of 'the way things seem' from the machine perspective. As Aleksander has argued elsewhere, and continues to argue here, the conscious machine must have the capacity for 'depiction', that is, a mechanistic equivalent of the Heideggerian Dasein or 'being there'. The authors then submit two models to their phenomenology tests: Shanahan's embodied concept of Baars' Global Workspace architecture, and Aleksander's own kernel architecture. They conclude that phenomenology can be considered at a number of different levels of mechanistic description, and that consideration of this sort can produce fruitful discussion of consciousness and provide practical ideas for how a synthetic phenomenology might lead to the design and development of new functional artefacts.

Synthetic phenomenology is an interesting concept, drawing together the enacted-unconscious, or 'phenomenal-consciousness' (Block, 1995), with depicted-consciousness or 'access-consciousness' (Block, 1995). There is great mettle in their three-fold attempt to (i) make computationally clear the relation of first-person phenomenal states to their world, (ii) explain how meaningful states arise in the absence of meaningful sensory input, and (iii) describe how a sensation of 'what to do next' arises in an agent; and, their enterprise is, for the most part, extremely successful. However, there are a couple of niggling elements, neither of them damning criticisms.

The first is probably a fairly minimal concern. It is the suggestion that there could be a 'perfect knowledge of the world' were it not for the weakness of our sensory transducers. Is this really a transducer problem? As active participants in the perception and organisation of our experience, it is more likely to be the result of the inevitable effect of perceptual interference, with perfect knowledge being something we hold only as a Platonic ideal. The second concerns their notion of 'depiction'. The axioms that underpin depiction do not present a picture of 'simple' phenomenal-consciousness, nor of phenomenal- and access-consciousness combined; if anything, depiction over-specifies itself in a more robust manner as a self-consciousness that requires not only being-there but also being-here or Fort-sein. The feeling of 'being the focus of an out there world' can only be conceived if there is a 'there' of which I am part as 'here'; it is this localisation in space which can provide the organism with its point of view. But maybe this is all to the good for the authors whose thesis might just provide them with more than they'd bargained for!

What would it mean for an artificially conscious system to make sense of something? In 'Sense as a "Translation" of Mental Content',

Andrea Lavazza addresses this point by looking for philosophical support for the phrase 'to make sense' of anything. Distinguishing *sense* from *meaning* and *sensation* he sees the concept as the instance where a set of mental objects translate from one to another so as to have a phenomenological overlay of there being no contradiction in this process. The paper mainly sets this idea in the context of existing concepts of models of mind, philosophical and machine-oriented. For example he argues that sense may exist in the mind of the non-Chinese manipulator of symbols in the Searle Chinese room if the operations of matching incoming symbols form a closed translatable set of intensions that are independent of the meaning of the Chinese symbols. He extends his notion to something that makes sense in a societal context and, to discuss operationalisation, refers to Lenat's Cyc system where 'common sense' primary assertions are stored and where the resulting concatenation of such concepts into further assertions that make sense is compatible with the notion of translation. A major section is devoted to James' concept of 'fringe', that controls the relationship between coherent thoughts in the putative stream. He concludes that *sense* does not correspond to *fringe,* leaving fringe as being part of sense, but not the other way round. There is much more in this paper making it possible to agree with the author's conclusion that the '*translation* as the basis of *sense'* notion can support further research.

Complex environments, says Salvatore Gaglio in 'Intelligent Artificial Systems', require complex organisms. Certain kinds of complex organisms, he suggests, require the ability to process symbols and assemble them into expressions, even though 'it is clear that if we introduce such expressions into a machine, it doesn't mean that it understands the sense of them at all' (p. 103). Beginning with Turing's imitation game — more popularly known as the Turing Test — Gaglio offers a tour of some of the highlights of the last fifty years in artificial intelligence, from first-order predicate logic to the search problem to heuristic shortcuts to the physical symbol system hypothesis, symbol grounding, and symbol semantics. Gaglio may be unfair to Turing in his account of the imitation game, suggesting (along familiar lines) that Turing was trying to provide a conclusive test for machine intelligence (rather than, say, offering a precursor to Dennett's intentional stance). But then Turing's paper has seen fifty years of continuous reinterpretation, the meaning assigned to it often having more to do with the needs of the moment than whatever Turing may have had in mind.

Likewise Gaglio's account of the history of AI is curiously focused toward what now is commonly known as Good Old-Fashioned AI, though he offers some discussion of neural networks by way of balance. His statement that the Church-Turing Thesis is 'not a theorem but a fact' (p. 100) is, perhaps, overstating the case. What may be the most interesting part of the paper is his discussion of Gärdenfors' notion of conceptual space as analogous to physical space: the idea that abstract concepts may have a kind of length, width and height of their own; and that considering concepts in this way may offer a tidy solution to the symbol grounding problem, by making it possible for 'all the symbols [to] find their meaning in the conceptual space that is inside the system itself' (p. 111).

The question Gaglio returns to again and again is, where does meaning come into the system? If the system is to be truly intelligent, it can't just come from the observer. The intended destination for all of this discussion is what Gaglio calls 'the self of the robot': the development, in an artefact, of a concept of self. Consciousness arises, he suggests, through the iterative collapse of a sequence of representational distinctions. It is unfortunate that this part of the paper is the most brief and leaves us with many tantalizing questions of how, precisely, the self fits in, what the collapses mean, and where the discussion might go next.

The chapter by Maurizio Cardaci, Antonella D'Amico and Barbara Caci is somewhat misleadingly titled 'The Social Cognitive Theory — A New Framework for Implementing Artificial Consciousness'. The first part of the title stems from the fact that the work is inspired by Bandura's (1986; 2001) social cognitive theory. Apart from this, however, somewhat surprisingly, the chapter is not at all about social cognition or behaviour. Instead the focus is on the role of different types of conscious processes in the regulation of individual motivated behaviour.

Following Bandura's view of 'triadic reciprocal determinism', the authors take *emergent interactive agency* to be crucial to consciousness and to be the result of the interaction of actions, personal cognitive/affective factors and environmental events. Core features of consciousness arising from emergent interactive agency are, according to this view, intentionality, forethought, self-reactiveness, and self-reflectiveness. While intentionality generates goal-oriented actions, forethought is about predicting the likely consequences of possible actions. Self-reactiveness is about self-monitoring and -correction, whereas self-reflectiveness is a meta-cognitive ability that allows an agent to examine its own thoughts and actions.

Based on these considerations, the authors implemented (in previous work) a robotic architecture that allows agents to generate plans of actions, based on initial expectancies, and compare expected with actually obtained results. The authors also experimented with internal and external locus of control models; in the former case the mood state (which affects execution speed and new plans) is updated based on how well expected and obtained results match, whereas in the latter case the update depends on a randomly generated value. As future work the authors discuss a more flexible architecture in which metacognition would be used to adapt the locus of control: in a predictable environment presumably an internal locus would be most useful, whereas in unpredictable environment an external locus of control would probably be preferable.

The main contribution of this work is that it makes use of Bandura's work on consciousness which otherwise has received only little, if any, attention from robotic/computational modellers of consciousness. The models discussed in this chapter, however, are too abstract and limited to really do Bandura's work justice, not least when it comes to the actual role of social and cultural factors.

Antonio Chella cheerfully informs us, at the outset of 'Towards Robot Conscious Perception', that 'In a word, new robotic agents must show some form of artificial consciousness' (p. 124). This will come as rather a shock to many roboticists who had hoped their business was refreshingly free of philosophical conundrums treated through the years on the pages of *JCS:* zombies and their relatives, such as (supposedly) qualia-lacking computers built of beer cans and string. Of course, the phrase 'In a word' here signifies that preceding it is a wordier, and perhaps more plausible, presentation of the claim that these 'new robotics agents must show some form of consciousness'. Well, in the preceding, we are informed that:

> A new generation of robotic agents, able to perceive and act in new and unstructured environments should be able to pay attention to the relevant entities in the environment, to choose its own goals and motivations, and to decide how to reach them. (p. 124.)

After reading this quote five times, with one's thinking cap on, one still fails to see why the decidedly consciousness-free robots in anyone's lab don't qualify for the title of 'new-generation of robotic agents', given the behaviours here cited as a measuring stick. And this is not even talking about research-grade robots, but rather about Legobots used in first-year undergraduate robot instruction: robots able to negotiate novel versions of the famous Wumpus World (amply described in Russell & Norvig, 2002), often used to challenge simple mobile robots.

The very same point can be made about the robot that Chella features in this chapter: viz., *Cicerobot*, a museum tour guide in Italy. For example, Chella tells us that this robot enjoys a 'conscious perception loop'. But the loop is diagrammed in Figure 8, and after study of this figure it is hard to see why the garden-variety dataflow shown there should be labelled with anything other than the straightforward phrase 'perception loop'.

The diagnosis of the chapter can be generalized: While the engineering described appears to be competent, prefacing rather standard engineering processes with loaded philosophical terms does not suffice to bestow upon the artefacts in question the properties associated with these terms, and most roboticists will simply ignore these terms anyway.

In 'A Rationale and Vision for Machine Consciousness in Complex Controllers', Ricardo Sanz, Ignacio Lopez and Julita Bermejo-Alonso declare that 'software intensive controllers are becoming too complex to be built by traditional software engineering methods.' Were this true, there is little question we would have on our hands a worrisome state of affairs — if for no other reason than that, at least as far as we can tell, the state of the art in formal verification of the behaviour of software would be classified by Sanz and co-authors as traditional. At any rate, leaving the consequences of the potential failure of traditional methods aside, is it in fact true that they *are* obsolete?

Sanz *et al.* certainly think so; they boldly state:

> We have reached the conclusion that the continuously increasing complexity make almost impossible the use of construction-time techniques because they do not scale and prove robust enough. (p. 143.)

But no arguments are provided in support of this claim. Software controllers are by definition implementations of functions that can be formally defined ahead of time (after all, these controllers are built because implementations of certain known-ahead-of-time functions are sought), and there seems to be no reason to believe that one cannot formally express the functions in declarative form, and prove that one's implementation coincides with what is needed, precisely. In fact, there have been unprecedented advances in this direction (e.g., see Arkoudas *et al.*, 2004).

Be that as it may, it's certainly quite interesting that the authors make what they call a 'business case' for conscious machines: the idea being that in light of the purported failure of traditional methods,

'conscious' software controllers are needed. (Sanz *et al.* augment this case with the claim that, from an evolutionary perspective, consciousness must be very, very valuable, but they seem to be unaware that some have looked at this from a radically different perspective: viz., that since creatures without consciousness, but our behavioural power, could have evolved, but didn't, we are faced with a profound mystery. See Bringsjord *et al.*, 2002.)

Unfortunately, the authors seem not to realize that if what they describe as a conscious controller (and, more generally, a conscious machine) is indeed conscious, then so are mundane logic-based systems in AI. For example, they present (in their Figure 2) an example of a conscious system with sensors, effectors, and a knowledge base, that operates in a loop as it interacts with the environment. But this system would seem to be almost a perfect match with the agent model presented in matching diagrams in Nilsson (1991) — and yet Nilsson doesn't in the least classify this agent as conscious.

For us to recognize a robot as conscious, suggest Owen Holland, Rob Knight and Richard Newcombe in 'The Role of the Self Process in Embodied Machine Consciousness', consciousness must be sufficiently analogous to human consciousness; and that in turn requires the robot to be embodied, at some level of abstraction, in the same manner as a human. CRONOS, touted as the first anthropo- mimetic robot, was designed from a textbook on human anatomy and looks the part.

An agent, be it natural or artificial, is, for the authors, an agent on a mission. For living organisms the primary mission, from an evolutionary standpoint, is reproduction. Simple organisms can achieve their mission through purely stimulus-response mechanisms. Flexibility is gained by allowing the agent to modify its behaviour based on aspects of its environment not immediately apparent through its senses: simple induction and deduction. To go beyond that, they say, requires the ability to go beyond experience: to imagine the world not as it is but as it might be. Now there is not just agent and environment, but, within the mind of the agent, a model of the agent and a model of the environment, which are put to use in simulation after simulation.

So the mind of the self-conscious agent is populated with representations, some of which are special because they are representations of the agent itself. Consciousness, the authors suggest, is simply an emergent property of those self representations. As the agent interacts with its environment, so the representation of the agent interacts with, and is conscious of, the agent's representation of its environment,

which it takes to be the real thing. The mission, here, is to achieve as close a fit as possible between the latter and the former.

CRONOS, the robot, meets SIMNOS, the representation of robot and environment built with software designed for the games industry. Though there is no claim of consciousness here — yet! — there is the suggestion that a more complete implementation of CRONOS and SIMNOS might well get there. Sceptics will not be convinced, but partisans of machine consciousness will find much to encourage them in what is in many ways a novel and thought-provoking approach.

In writing 'From Artificial Intelligence to Artificial Consciousness', Riccardo Manzotti is to be commended for recognizing a number of distinctions often glossed over by non-philosophers. For example, the distinction between shallow senses of consciousness and full-blown subjective consciousness is made. In the case of the latter phenomenon, the problem, from the standpoint of robotics and computation, is how to express, in rigorous, third-person terms, that which it is like to (say) taste deep dark chocolate ice cream. The problem is expressed, and argued to be unsolvable, in Bringsjord (1995; 1999).

Manzotti claims to provide a solution to the problem in this very chapter. Were such a solution to in fact be provided, the chapter would soon enough come to be regarded as seminal. So, what is the solution that is supposedly supplied?

We read:

> As soon as we drop the belief in a world of things existing autonomously and as soon as we conceive the world as made of processes extended in time and space, experience (and thus consciousness) does not need to be located in a special domain (or to require the emergence of something new) — experience is identical with those processes that make up our behavioural story. (p. 181.)

Manzotti encapsulates his bold move by proclaiming: 'The traditional problems of phenomenal consciousness vanish once an externalist and process-based standpoint is adopted' (p. 183).

Unfortunately, this move, even under the assumption that the problems in question do indeed vanish, is anaemic. The reason is simple: Philosophy doesn't work by legislation, but rather by argumentation. If the former technique were viable, then the sub-fields of philosophy would be rather easier to manage. In ethics, we could settle the problem of abortion once and for all by having everybody drop the belief that abortion is morally wrong; in philosophy of religion we could settle the main issue once and for all by having everybody drop the belief

that God exists, despite arguments offered by Anselm, Descartes, and Gödel; and so on for the other sub-areas.

Finally, Manzotti must face up to a second problem: Didn't he write the chapter in question? If his externalist/process-based standpoint is affirmed, then human persons aren't determinate entities who deserve credit for autonomously doing anything. His view thus seems self-refuting: We can only take it seriously if he articulates compelling arguments for it, but the view itself entails that 'he' can't 'do' anything.

In his earlier paper 'Internal Robotics', Domenico Parisi (2004) investigated how robots could achieve greater fitness and flexibility by not just reacting to the external environment in which they were embedded but also to some simulated internal bodily dynamics. By contrast, his article in this volume, 'Mental Robotics' suggests that yet more flexible robots need the ability to self-trigger internal representations and that this is key to their having a mental life. Representations, Parisi argues, are formed by organisms and robots as ways of producing actions in the face of diverse sensory information. Some organisms develop the ability to take this a stage further as a way of dealing with entirely absent ambient information, i.e., they come to trigger their own representations. Representations are needed because we cannot always rely on ambient environmental information to clearly tell us what to do.

Properly 'mental' images, Parisi argues, are those that are not caused directly by environmental information but are self-generated internally. Robots that can use such self-triggering begin to have a mental life. There are a variety of forms of mental life that rely on these images such as planning, recollection, dreaming and hallucination. In this Parisi agrees with others who have taken the simulation approach to consciousness (Hesslow, 2002) and to representation (Clark & Grush, 1999). Parisi holds that there is a special role played by self-generated linguistic episodes. This is because the mental images of words can provide advantages of economy over more fully elaborated mental images. From this he makes a case for the special properties of internalised language in mental life; something not all simulation theorists agree with (although compare Clowes, 2007, for a related account)

In all, Parisi proposes an ambitious programme for understanding mental life through building robots that use mental images in a variety of scenarios and in order to illustrate diverse mental functions. It will be interesting to see if this approach can also explain how such mental images are integrated in an overall presentation of the world; something that would also seem to be required for an account of conscious mental life.

As their title ('An Account of Consciousness from the Synergetics and Quantum Field Theory Perspectives') suggests, Alberto Faro and Daniela Giordano attempt to account for consciousness by combining elements from dynamics and quantum field theory. In attempting to go beyond Haken's theory of synergetics, which they feel is more suited to the description of unconscious and weakly conscious mental states, they propose a new theory, which they term FRT or 'The Framing and Reframing theory', inspired by Ervin Goffman's 1974 Frame Theory. They aim to explain consciousness as the process of observing ourselves adapt, mainly through learning.

In order to do so the authors define two spaces, an 'activity space', which is similar to Haken's synergetic space, containing organized activity patterns; and an ontological space, in which the patterns are classified according to their similarity. The interaction between the two spaces is meant to provide us with systems that can observe themselves adapt and undergo more efficient reframing by recalibrating the dynamic control parameters that enable Haken's systems to respond to a given context by activating a specific behavioural class.

While the authors believe that FRT is sufficient to explain consciousness (in the way that they define it) they also feel that the biggest problem of this abstract architecture is that it is not realized in classical brain theories. So they are forced to appeal to an extraordinary theory like Vitiello's 'dissipative quantum brain dynamics'. Yet we already have well established classical theories like Reverse Hierarchical Theory, or RHT (Borenstein and Ullman, 2002; Hochstein and Ahissar, 2002), that describe the interaction between areas V1 and V4 in the visual cortex in terms of reciprocal causation resulting in appropriate local frames; further, as Borenstein and Ullman note, this is just a small part of a bigger system with different modes of top-down causation.

Another reason the authors use to justify resorting to quantum field theoretic models of the brain is that the brain harbours non-local correlations that cannot be explained by classical physics. Despite Vitiello and Freeman's claims, there seem to be no data establishing that conclusion. However if it turns out to be true that any artefact, let alone the brain, could reliably use the infinite degeneracy of the vacuum ground state to store and retrieve information, that would be an incredible breakthrough. In the meantime not only do theories like RHT or even Baars' Global Workspace theory avoid the pitfall of

environmental decoherence that plagues quantum mechanical theories of the brain, they are supported by the biological data.

In 'The Crucial Role of Haptic Perception: Consciousness as the Emergent Property of the Interaction between Brain Body and Environment', Piero Morasso, known to many for his work on self-organising neural systems, extends the term 'haptic' beyond just meaning 'touch'. He lets it encompass the motor support that is necessary to touch something, the discovery through touch of what things are and how the integration of these processes goes towards the creation of a sense of a bodily *self*. There are some interesting thoughts on the role of haptic consciousness in transferring from pre-natal thumb sucking to post-natal breast-suckling and eventual human sexual experience. Piero Morasso in presenting this paper at a workshop at Agrigento, held at the Archbishop's palace in his very presence, argued as he does in his paper: 'The official wisdom is that sex is a strictly private matter whose only social acknowledged purpose is reproduction... A distinguishing human fact is to be able to completely uncouple sex from reproduction and use it as an expression of human interaction.' The Archbishop's comment is not recorded.

Morasso's thesis is that the way the brain integrates action and the haptic sense places a sensation of self at the point where the organism interacts with the world. This is a key feature of the mechanism for creating a part of 'self'. For example, using a screwdriver sometimes makes it seem that the world is 'felt' at the tip of the screwdriver. Much evidence is drawn from phantom limb experiments that provide a clue to the creation of consciousness through the interaction between brain, body and, importantly, embodiment in an environment. Consciousness of the phantom limb is so strongly adapted that it points to a fundamental way in which sensing of location through touch is important in being conscious of a solid world. There are clear parallels here with the notions of visual sensory-motor contingency ideas of O'Regan and Noë. All this, argues Morasso, is not only revealing in the formation of a bodily self, but also provides a better approach to therapy in the case of limb loss. Engagingly written, this paper raises questions in the context of artificial consciousness that are bound to be debated further.

In a witty and perceptive end-piece ('The Ensemble and the Single Mind'), Peter Farleigh critiques the functionalism that lies at the philosophical depths of the Artificial consciousness project. According to David Chalmers' principle of organizational invariance, highlighted by Farleigh, the same experience may emerge from each of two systems very different in physical makeup (e.g. brain vs. silicon)

provided 'the abstract pattern of causal interaction between components' is identical (the discussion goes into much more detail). Farleigh's chief aim is to put pressure on this notion of 'sameness of pattern of causal interaction'. He asks us to imagine having the causal links between a pain-receptor in one's finger and the nerve severed. A new connection is made via an external mechanism that exactly reproduces the normal timing between finger and neural pathway. If I now catch my finger in a door I still feel pain, but the cause of the pain isn't the damage to the finger, but the external stimulation device.

Farleigh then asks us to imagine an entire brain where interneuronal links are severed, and comprehensively replaced by myriad external stimulators devilishly timed to exactly ape the normal endogenous causal patterns of the brain. Would this dissociated version of me be experiencing the same qualia that I do? Farleigh argues that there are problems with both negative and positive answers to this question. His conclusion is that the simple-minded notion of causality assumed by functionalists is not able to do justice to our intuitions about how consciousness causally relates to the brain. This is an ingenious and subtle paper which, while it won't give artificial consciousness practitioners too many sleepless nights, may give pause to those who think the theoretical issues underlying the practice will be ironed out with relative ease.

References

Arkoudas, K., Zee, K., Kuncak, V. & Rinard, M. (2004), 'Verifying a file system implementation,' *Proceedings of the 2004 International Conference on Formal Engineering Methods* (ICFEM), Volume 3308, Seattle, WA, November, pp. 373–90.

Bandura, S. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory* (Englewood Cliffs, NJ: Prentice-Hall).

Bandura, S. (2001), 'Social cognitive theory: An agentic perspective', *Annual Review of Psychology*, **52**, pp. 1–26.

Block, N. (1995), 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences*, **18**, pp. 227–47.

Borenstein, E. and Ullman, S. (2002), 'Class-specific, top-down segmentation', in *Proceedings of the 7th European Conference on Computer Vision-Part II* (May 28–31, 2002), pp 109–22.

Bringsjord, S. (1995), 'In defence of impenetrable zombies', *Journal of Consciousness Studies*, **2**(4), pp. 348–51.

Bringsjord, S. (1999), 'The zombie attack on the computational conception of mind', *Philosophy and Phenomenological Research*, **59** (1), pp. 41–69.

Bringsjord, S., Noel, R. & Ferrucci, D. (2002), 'Why did evolution engineer consciousness?', in Fetzer, J. and Mulhauser, G., eds., *Evolving Consciousness* (San Francisco, CA: Benjamin Cummings), pp. 111–38.

Clark, A. & Grush, R. (1999), 'Towards a cognitive robotics', *Adaptive Behavior*, **7** (1), pp 5–16.

Clowes, R.W. (2007), 'A self-regulation model of inner speech and its role in the organisation of human conscious experience', *Journal of Consciousness Studies*, **14** (7), pp. 59–71.
Forman, R.K.C. (1999), 'What does mysticism have to teach us about consciousness?', in Gallagher, S. & Shear, J. (eds) *Models of The Self* (Exeter: Imprint Academic), pp 361–78.
Hesslow, G. (2002), 'Conscious thought as simulation of behaviour and perception', *Trends In Cognitive Sciences*, **6**(6), pp. 242–7.
Hochstein, S. and Ahissar, M. (2002), 'View from the top: Hierarchies and reverse hierarchies in the visual system', *J. Neuron*, **36** (5), pp 791–804.
Nilsson, N. (1991), 'Logic and artificial intelligence', *Artificial Intelligence* **47**, pp. 31–56.
Parisi, D. (2004), 'Internal robotics', *Connection Science*, **16**(4), pp 325–38.
Rushworth, M.F.S., Nixon, P.D., Renowden, S., Wade, D.T., and Passingham, R. E. (1997), 'The left parietal cortex and motor attention', *Neuropsychologia*, **35** (9), pp. 1261–73.
Russell, S. and Norvig, P. (2002), *Artificial Intelligence: A Modern Approach* (Upper Saddle River, NJ: Prentice Hall).
Vogel, E.K., Luck, S.J. & Shapiro, K.L. (1998), 'Electrophysiological evidence for a post-perceptual locus of suppression during the attentional blink', *Journal of Experimental Psychology: Human Perception and Performance*, **24** (6), pp. 1656–74.

Book Reviews

John R. Searle
Freedom & Neurobiology: Reflections on Free Will, Language, and Political Power
Columbia University Press, 2007, 113 pp.
ISBN: 978-0231137522

Reviewed by Henry Stapp

This short book is ideal for readers looking for a brief, clearly articulated account, by one of the world's foremost philosophers, of his opinions on a basic philosophical problem of our time — the problem of free will. The question is whether every action that you make is pre-determined by a causal process completely expressible in terms of what Searle calls 'mindless, meaningless, unfree, nonrational, brute physical particles' (p. 5). He asserts that

> We now have a reasonably well-established conception of the basic structure of the universe. ... We understand that the universe consists entirely of particles (or whatever entities the ultimately true physics arrives at), and these exist in fields of force and are typically organized into systems. On our earth carbon-based systems made of molecules that also contain a lot of hydrogen, nitrogen and oxygen have provided the substrate of human, animal, and plant evolution. These and other such facts about the basic structure of the universe, I will call, for short, the 'basic facts'. The most important sets of basic facts are given by the atomic theory of matter and the evolutionary theory of biology (p. 4).

That statement identifies the basis of Searle's approach: We human beings are biological systems made of atoms and molecules, and our understanding of ourselves should therefore emerge from an analysis of our understandings of our biological structures, which rest in turn on the atomic theory of matter.

Searle notes that his approach rests also on an important difference between what is possible in philosophy today and what was feasible in