Next-generation approaches to machine consciousness

Ron Chrisley[*]
ronc@sussex.ac.uk

Rob Clowes[*]
robertc@sussex.ac.uk

Steve Torrance[*†]
stevet@sussex.ac.uk

[*] Centre for Research in Cognitive Science
University of Sussex, Falmer, UK

[†] Institute for Social and Health Research
Middlesex University, Enfield, UK

Abstract

A spate of recent international workshops have demonstrated that machine consciousness is a swiftly emerging field of international presence. Independently, there have been several new developments in cognitive science and consciousness studies concerning the nature of experience and how it may best be investigated. Synthesizing results from embodied AI, phenomenology and hermeneutics in Philosophy, Neuroscience and enactive Psychology (among others), new paradigms for research into natural consciousness that transcend the limited behavioural/cognitive or neural/functional oppositions are being proposed and tested, with encouraging results. This paper gives an overview of some work that attempts to entwine these two strands to see how they might be of mutual benefit to each other.

1 Introduction

The goals of the field of Machine Consciousness are: 1) to create artefacts that have mental characteristics typically associated with consciousness (such as awareness, self-awareness, emotion and affect, experience, phenomenal states, imagination etc.); and 2) to model these aspects of natural systems in embodied models (e.g., robots). Machine consciousness symposia in Cold Spring Harbor (2001), Skövde (2001), Memphis (2002), Birmingham (2003), Turin (2003) and Antwerp (2004) have demonstrated that this is a swiftly emerging field of international presence.

Independently, there have been several new developments in cognitive science and consciousness studies concerning the nature of experience and how it may best be investigated. Synthesizing results from embodied AI, phenomenology and hermeneutics in Philosophy, Neuroscience and enactive Psychology (among others), new paradigms for research into natural consciousness that transcend the limited behavioural/cognitive or neural/functional oppositions are being proposed and tested, with encouraging results.

Next-generation approaches to machine consciousness attempt to entwine these two strands to see how they might be of mutual benefit to each other. A guiding principle behind this union is that advances in consciousness research can guide efforts into building conscious systems. But equally, there is a belief that the converse is true: The insights gained from attempting to build embodied, experiencing agents can provide important feedback to the various disciplines of consciousness studies. At the very least, the difficulties we encounter in our attempts to build systems which instantiate or model cognitive phenomena can point out where our current theories are incomplete, inadequate or incorrect.

The symposium entitled “Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment” (part of AISB 2005) brings together active researchers in this area and provides a forum in which their work may be compared, contrasted, evaluated and discussed. This paper uses the work being presented at that symposium as a framework around which to organise a survey of some of the key strands in contemporary work in machine consciousness.

The term “next-generation” may be something of a misnomer, since there is no clear consensus as to what constituted “first-generation” approaches to machine consciousness; certainly no attempt will be made here to provide a scholarly account of earlier approaches in this area. Nevertheless, we feel that the application of recent advances in our understanding of consciousness to the construction of working systems constitutes a major milestone on the way to achieving machine consciousness.

After a general discussion of the enterprise of machine consciousness in section 2, sections 3-10 examine what we believe to be eight key areas of consciousness studies that are best placed to help make progress on machine consciousness: work on imagination, emotion (and feelings of emotion), development and self-creation, enactivism, heterophenomenology, synthetic phenomenology, second-person approaches and neurophenomenology, and ethical and legal issues.

2 Machine consciousness: The very idea

As Torrance (2005, this volume) points out, one can revisit Searle's (1980) old distinction between weak and strong AI and similarly differentiate between weak and strong forms of machine consciousness. The first attempts merely to model conscious states in an artificial system, without any ambitions of actually replicating consciousness in that system. Strong machine consciousness goes further, and seeks to create artificial systems with experiential states themselves.

One might think it folly to engage in machine consciousness research (especially of the strong variety), given the opposition that confronts it on both sides. On the one hand, there are some popular arguments (e.g. Jackson, 1982) against physicalist accounts of consciousness, which claim that some form of dualism is the case. On the other hand, there are the arch-physicalists, who define consciousness in such a bio-centric manner that no non-biological system, such as the ones with which the field of machine consciousness typically concerns itself, could ever be conscious. Machine consciousness appears trapped between these two extremes; surely one or the other of them must be correct, and yet both rule out the possibility of conscious machines.

Although this is not the proper place for a full reply, a quick response can be made, since it works equally well against both lines of attack. Many in the machine consciousness community take what has been termed (since the Birmingham meeting) an "engineering" approach. Rather than claiming to have a solution to the "other minds" problem that would let them know definitively whether or not this or that artefact is or could be conscious, these researchers are more pragmatic. Modifying a criterion from the field of artificial intelligence, they will have considered their goal accomplished if they design and construct a system that does the kinds of things that, when done by a human, requires consciousness. It is sufficient for them to produce a system that behaves in such a way that, if it were an organism, we would assume that it is aware.

Many researchers in the area find the axioms provided by Aleksander and Dunmall (2003) to be of assistance in guiding their research. More generally, it is thought that the goal of the field will have been achieved if one can impart to a robot some combination of features, possibly including some of the following:

- Autonomy
- Adaptivity/advanced learning capacities
- Emotion/affect
- Responsibility (or being something to which we are responsible)
- Intelligence
- Authenticity (own world view and goals)
- Ability to integrate information from different sources/modalities
- Vivid/meaningful sensation/perception
- Ability to act in the world
- Ability to simulate/imagine/plan
- Ability to represent its own states
- Attentional capacities
- A belief that it is conscious/an ability to give phenomonological reports

Certainly most in the field would consider their primary goal achieved if they could build a system which had all of these features, even if philosophical doubts as to whether the system is "really" conscious might remain.

But even on that point, there is room for optimism. If Sloman and Chrisley (2003) are right, then current philosophical puzzles concerning how we could ever know a machine is conscious (a product of the apparent possibility of "zombies" (Chalmers, 1996)) might be features of our current, inchoate concept of consciousness. Perhaps we can't get to the successor concepts of consciousness that will solve these problems through armchair theorizing alone. But if we design, build and interact with artefacts with some of the properties listed above, that might be enough to cause our concept of consciousness to evolve until we see that no, it isn't possible for a system to have *this* architecture and not be aware. Zombies may seem possible now, but the kind of research surveyed in this paper might someday reveal that they actually are not possible

2 Imagination

A key finding of the Birmingham and Turin meetings was the existence of a common theme in much of the recent work in machine consciousness: The imagination or simulation approach. The basic idea is of an ability to predict, given the current sensory input, the future sensory input one would receive if one were to make a particular motor response. If this predicted sensory input is used as the "current" sensory input for an iterated application of the predictive process, one can anticipate the sensory input one would receive if one made a second motor response after the first, and so on. This allows entire sequences of behaviours, with the corresponding sequences of sensations that would occur during that behavioural sequence, to be "imagined".

The idea of using a simple recurrent network to give a robot this kind of imaginative capacity is not particularly new (see, e.g., Chrisley, 1990). But from the start it was acknowledged that imaginative capacities that dealt only in the lowest levels of sensory and motor encodings would be extremely limited. The work of Stening, Jacobsson and Ziemke (2005, this volume) is therefore a welcome development in this area of research. Not only do they incorporate an abstraction mechanism that allows their robot to imagine at a higher level of "conceptualisation" than the lowest sensory and motor levels; they also provide an inversion mechanism so that the imagined abstract states can be converted into expected sensory-motor states. A future extension of this work might be to have both low-level and abstract-level imaginative capacities working simultaneously, so that expected low-level sensations can be fed into the abstraction mechanism to yield a second route to abstract expectations. Actual abstract expectations might be some kind of average between the "abstract-then-imagine" expectations and the "imagine then abstract" expectations outlined here.

Shanahan (2005, this volume) illustrates the imagination approach very well. He reports on a new kind of design for robot architectures that incorporates two linked action-generation systems, a first-order reactive system and a higher-order one which introduces off-line 'imaginative' rehearsals of action alternatives in a way that modifies the saliency levels made available to the first-order system. The resulting architecture incorporates various key features of mammalian brains. The function of imaginative rehearsal plays a key role in the model of consciousness offered by Shanahan. The model provides, in his view, a useful approximation to the role played by consciousness in real agents.

It is hard to say exactly what it is about imaginative processes that makes some researchers take imagination to be essential to consciousness. For some, such as Haikonen (2005, this volume) and Stening et al. (2005, this volume, following Hesslow (1994)), consciousness consists in having an inner life or inner world, and it seems more plausible to say an artificial system has such if one can identify states of the system that are of the same format as perceptual states, but which correspond to anticipated rather than actual sensory input. The imagination approach, with its extension of perceptual processes to cognition as a whole can be seen as a new kind of empiricism. Yet, as Haikonen points out, imagination allows one to transcend perception, in that one's behaviour may sometimes be driven by internal (albeit pseudo-perceptual) processes rather than the current sensory input. A striking feature of his model is its attempt to go beyond the simplest models of artificial imagination, by integrating it with elements such as attention and decision-making.

Stuart (2005, this volume) also addresses the relation of imagination to consciousness, but in a rather different way. She suggests that Kant's transcendental philosophy prefigures a variety of recent studies of artificial agency and consciousness – particularly the work of Cotterill, Sloman, Aleksander, Bowling and others. Her focus is on Kant's treatment of the problem of how it is we can take the order of our experiences as belonging to a unified 'I' – a precursor of the contemporary binding problem. Kant's solution is complex but, as Stuart shows, a central strand appeals to the imagination, specifically the cognitive or productive imagination that (working with the senses and the understanding) enables us to treat each of our experiences as modifications of he same mind; as linked in consecutive, associative patterns; and as similar or different from preceding experiences. Kant distinguishes between productive and reproductive imagination: the first is essential for any thought and necessary for the constitution of self-consciousness; the second is the ability to bring to mind things that are not wholly present. No doubt both types of imagination are required within an adequate model of consciousness.

One of the first intended applications of imagination in robots was planning (Chrisley, 1990; Stein, 1995). Chella, Frixione, and Gaglio (2005, this volume) combine this idea with a linguistic abstraction capability to allow for grounded planning for linguistically-specified goals. It seems that their system could be extended to also allow for linguistically-specified environmental information to play a role in the planning process.

3 Emotion (& feelings of emotion)

In recent years the seeming antithetical study of emotions in machine systems has started to be treated seriously (Picard, 1997). In work on machine consciousness, Aleksander, Lahnstein, & Lee argue, one should be "suspicious of the consciousness of a machine were it not to have mechanisms that play the role of emotions" (2005, this volume). They maintain that valenced evaluation of the state of the organism, both actual and projected, are central to the long-term viability of, and the development of the capacities of, the organism. Some researchers in machine consciousness seek to develop this idea with reference to the ideas of Antonio Damasio.

According to Damasio, emotion is not only central to reasoning (Damasio, 1994) but to the generation of what he calls core consciousness (Damasio, 2000). On Damasio's account, core consciousness emerges for an organism as it becomes able to detect that its core body state has been changed by some incoming stimulus. The reactive component of the organism's neural representation of such a stimulus is conceptualized as an emotion. Bosse, Jonker, & Treur (2005, this volume) formalize this theory into a model expressing temporal and causal dependencies using their Temporal Trace Language (Jonker & Treur, 2002). Their formal model also predicts the possibility of "false core consciousness", where an effect is attributed to the wrong body stimulus.

Aleksander et al. (2005, this volume) build upon Damasio's model in order to understand a key point of discussion in the (natural) consciousness literature, that is, accounting for the reality or otherwise of "the will". Since the publication of Libet et al's (1983) finding that a neo-cortical readiness potential seemed to precede the ability of a subject to attest to willed action, the folk conception of volitional action has been called into question. One radical sceptic (Wegner, 2002) has recently argued that Libet's findings should be interpreted as showing that an unconscious cortical event controls both the "willed" action itself and the conscious sensation of control. By this reasoning, "the will" as currently conceptualised by the folk is simply an epiphenomenal shadow or illusion. Aleksander et al. (2005, this volume) instead see Libet's experiment as having taken volition out of its normal emotional envelope. By developing a model of how such volitions are typically generated within a framework of ongoing affective evaluations, the authors show that Libet's paradigm is actually an atypical example of willing where the will is relegated to *when* and not *what.* If volition is examined in its typical and proper emotional context, they argue, it approximates much more closely the way it is seen by the folk.

No doubt the next round of machine consciousness research will pay more attention to emotion and affect. In a rather plausible Humean way, Haikonen (2005, this volume) contends that an "emotional value system", or at least some affective distinction between pleasant and unpleasant, is required for decision making (in his case, via an imaginative system). Stening, Jacobson and Ziemke (2005, this volume) make a similar point, noting that future development of their work should allow the robot's needs to play a role in motivating and guiding its action, abstraction and imagination.

4 Development and self-creation

Development has for a long time been argued to be a crucial component in the understanding of consciousness. Vygotsky (1986), for one, pointed out this link by attempting to show how consciousness depended upon intersubjective social interactions. Although the connections between these areas remain largely unaddressed, work on the development of intersubjectivity (Trevarthen, 1994) may point the way forward; indeed, this work is starting to be taken very seriously in the related field of epigenetic robotics.

Of course, development appears to depend not just on external scaffolds but also on the developing bodily and situational substrate, and it is the attempt to understand the relation between these that has been fundamental to the concerns of epigenetic roboticists. A series of annual conferences in this field (starting in 2001) has focused on the question of how a robotic system, through extensive interaction with its environment, can transform itself from a being a purely reactive system into a fully intentional one. One idea is that this can happen only if an agent undergoes a prolonged developmental period (Zlatev, 1999). Central issues in the development of agents are the distinctions between self and other, body and environment, sense and action.

Such questions should also be germane to work on machine consciousness, not least because some would be unwilling to treat as conscious any system that was incapable of undergoing a process of ontogenesis – although this is controversial. This of course throws open the question of what forms of development might evince consciousness. One possible focus is the development of self.

As having a self – generally one to a body – seems to be typical of the type of consciousness we best know, i.e. our own, systems that attempt to explore the development of self should be of special interest. Many accounts of the self stress that a sense of self is not pre-given to an agent or merely represented internally, but is developed and maintained out of the sensorimotor flows in which the agent participates (Butterworth, 1998). Although there is considerable controversy over what is pre-given and what is developed (see for instance Gallagher & Meltzoff, 1996), the flexibility of the nature of the body image in higher animals now has extensive experimental demonstration (Ramachandran & Blakeslee, 1998). Iizuka & Ikegami (2005, this volume) argue that "body image and ownership" – concepts that seem closely related to the idea of the self – cannot be derived from static sense data alone. They argue that the self depends on and must be understood in terms of the emergence of the distinction between self and world. Inspired by Gibson's (1962) cookie-cutter experiment, they discuss a simulated agent that develops distinctions between 'sensor' and 'motor' through interactions with its world. They argue that the sort of active perception system here developed can help us understand the emergence of a self in a way which is precluded by the prior specification of sense and motor.

5 Enactive approaches

Enactive approaches to cognitive science have become popular of late. Enactivism was first formulated as an attempt to move beyond cognitive science methods dominated by cognitivist and connectionist paradigms (Varela, Thompson and Rosch, 1991). Strong emphasis was laid on linking cognitive science with insights from the hermeneutic phenomenology of Husserl and Merleau-Ponty, and in particular stressing the sensorimotor embodiment of an experiencing agent in a world, "enacting" that world and her own self in relation to the world.

There are a number of strands to the enactive approach. One focuses on perceptual experience, arguing that it consists in the exercise of the mastery of sensorimotor contingencies, and that awareness consists in the application of this mastery to a reasoning process (e.g. O'Regan and Noë, 2001; Noë, 2002). This view contrasts with the conventional view of perceptual awareness, according to which experience consists in sensory inputs generating internal, neurally-encoded representations of an external environment. For the enactive approach, perceptual consciousness has relatively little to do with intenal structures in the brain, and much more to do with ongoing sensorimotor and bodily interactions with the environment.

In this respect the enactive approach contrasts quite strongly with at least some variants of imagination-based approaches to modelling consciousness – for example that of Shanahan, who puts considerable emphasis on providing an architecture that reproduces detailed structures of the brain. Another prominently neurophysiologically-based approach to consciousness which, however, also lays great stress on embodiment and sensorimotor fusion in a way that is close to the enactive approach, is to be found in Cotterill (1998). Stuart (2005, this volume) considers Cotterill's approach in some detail, putting it into the context of Kant's debate with Hume over the nature of the unity of self-consciousness. She points out that an adequate account of the unity of the experiencing and active "I" must necessarily be strongly embodied, and thus her approach is also close to that of the enactive viewpoint.

Both Haikonen (2005, this volume) and Ikegami (2005, this volume) take there to be a fundamental connection between consciousness and enactive perception, at least in the Gibsonian sense that perceptual experience is not the passive reception of sensory inputs, but an exploratory interplay between the internal states of the agent and the external world. Haikonen points out that if perceptual experience consists in active exploration of sensory-motor contingencies, then it makes sense that imaginary or inner experience consists in the exploration of the interdependence between hypothetical motor commands and the anticipated sensory states which result. Ikegami's focus, however, is on exploration in the real world rather than in some inner simulation. He attempts to model this with an agent whose chaotic dynamics are such that the agent, he says, is not simply responding to the stimuli at any one time, but to a more abstract entity: the time structure of the stimuli. However, it is not yet clear whether such a dynamics has the property which Ikegami seeks: that of being able to specify what is characteristic of conscious states (or even living states. For Ikegami, life seems to be a prerequisite for consciousness, a contentious view in the machine consciousness community).

If the first-strand enactivists are right, then perceptual experience consists in the exercise of the mastery of sensory-motor contingencies, and that awareness consists in the application of this mastery to a reasoning process. In that case, the central goals for machine consciousness research would be a) establishing clear criteria for when a robotic system possesses such mastery and b) building robots which meet these criteria in a way which allows said mastery to play a crucial role in their deliberations.

A second strand in enactivism goes back to Varela's earlier work, with Maturana, on autopoiesis (Maturana and Varela, 1987). Autopoiesis is the process whereby an organism continually recreates itself in relation to its environment, through a process of internal self-regulation and the maintenance of a semi-permeable boundary through which matter or energy can be exchanged. There has been some interesting recent philosophical work exploring the implications of autopoiesis in ways that help to understand the nature of consciousness. Torrance (2005, this volume) discusses the significance of some of this work (e.g. Hanna and Thompson, 2003; Weber and Varela, 2002) in offering a new departure for machine consciousness. He considers an impasse over the 'explanatory gap' which is seen by many as blocking physicalistic attempts to explain the nature of phenomenal consciousness Torrance suggests that there is a defective concept of consciousness underlying this gap – 'thin phenomenality' as he calls it – which is also shared by many of those who think the gap can be bridged, including many machine consciousness researchers. An alternative, 'thick' conception of phenomenality is proposed, which takes ideas of autopoiesis, lived embodiment and other related ideas as its starting point.

6 Heterophenomenology

It seems undeniable that phenomenological reports are a valuable source of data concerning consciousness, and yet a scientific theory of consciousness must be sensitive to the possibility that subjects may be mistaken in their sincere avowals concerning experience. Dennett (1991; 2003) outlines a way to avoid the pitfalls of naïve or folk conceptions of consciousness without discounting phenomenological reports altogether: Heterophenomenology. Adopting this methodology with respect to machine consciousness seems promising, but poses difficult questions. For example, since linguistic phenomenological reports play such an important role in this approach, what kinds of communicative or linguistic abilities need a robot possess in order to allow the direct application of heterophenomenology?

Modelling and robotic work such as the Adaptive Language Games project of Steels (1998) and his collaborators has provided one way into understanding how mechanisms for grounding communicative symbols in perceptual abilities might be effected. In a recent extension to this work Steels (2003) has argued that a variation on the adaptive language games model can be used to help understand the inner-voice which is thought to be the constant accompaniment of much human conscious thinking (Hurlburt, 1990). In Steels' model, agents pre-check the interpretability of a putative sentence by feeding back the output from their production systems into their interpretation systems. It is argued that this "re-entrance" of linguistic information where an agent checks an utterance by projecting it back onto itself, explains the functional system underlying the phenomenology of inner speech.

Other work on linking cognitive and linguistic functions can be found in Sugita & Tani's (2002) report on their work with a mobile robot, where the robot comes to associate action categories and linguistic labels. Chella, Frixione, & Gaglio (2005, this volume) report on their work on Cicerobot, where a comparable approach is taken but on a more sophisticated mobile robot. In this research the authors have built a robot capable of vision and action which has an architecture based on linguistic, conceptual and sub-conceptual capacities. Cicerobot's architecture, however, is based on a three-layer model composed of a "subconceptual area… concerned with the processing of data coming from the robot sensors…, [a] linguistic area of representation and processing… based on a semantic network formalism… [and a] conceptual area intermediate between the subconceptual and the linguistic areas." Cicerobot's linguistic and subconceptual areas are used in behavioural planning and affective evaluations, and these different representational levels are mediated by the 'conceptual area'. The authors make use of conceptual space theory (Gärdenfors, 2000) to "provide a principled way for relating high level, linguistic formalisms with low level, unstructured representation of data." It would be interesting to see how this work might be developed to support the sort of phenomenological reports required for heterophenomenology.

However, the generation of narratives which would serve the role assigned for them by Dennett would seem to require the involvement of language in the ongoing activity of the agent in a way which would need to go somewhere beyond the labelling by the agent of its environment or even a role in planning (Clowes, 2003). Whether this work can provide a sufficient underpinning for machine heterophenomenology remains to be shown, but we are starting to have a better range of possible scenarios to consider.

Another direction in which to pursue this approach would be to try to make sense of infra-linguistic forms of phenomenological "reports". It seems possible at least in principle that a system incapable of using language might nevertheless attempt to represent its internal states *as* phenomenological states. Indeed, one might think such self-modelling might be a crucial component in explaining even the phenomenological reports of linguistic creatures. As Sloman and Chrisley (2003) point out, explaining why a system finds it useful to think (or speak) of itself as having qualia might go a long way to explaining the having of qualia itself.

7 Synthetic phenomenology

A science of consciousness, be it of natural or artificial agents, requires some ability to specify and refer to subjective, fine-grained experiential states, which, by their very nature, elude linguistic expression. One idea is that the states of artefacts-in-an-environment might themselves serve as ways of specifying the conscious states that they embody

(Chrisley, 1995). The sub-field of synthetic phenomenology aims to investigate this idea by, e.g., constructing means of visualizing or otherwise communicating the (actual, or modelled) experiential states of robots.

It has been known for some time that capturing the spatial content of experience is particularly problematic. Previous attempts to do so have simplistically plotted the robot's actual or imagined sensations on a map of objective space, even when the robot had no understanding of the connection between the spatial content of the sensors and movement, and even when the non-objective, non-systematic spatial representations of the robot were explicitly the topic of investigation (e.g., Chrisley, 1993). Another difficult area for synthetic phenomenology, discussed at the Antwerp meeting, is the specification of the content of experience which is more abstract or conceptualized than the lowest level of sensory and motor signals.

Stening et al. (2005, this volume) manage to make headway on both problems with a single solution: de-abstraction, or "inversion". Their initial representations of the abstract aspects of their robot's experience suffer from the usual problems: They are located on a map of objective space, and their forms (e.g., their gray-scale ordering, their circular shape) do not carry any content for us that is related to the contents of the robot which contain those abstractions. But their later representations of the robot's experience do much better: By inverting the sequence of abstractions into sensorimotor combinations, they are able to reveal the spatial relational structure of the robot's experience. The inversion, by reducing abstractions back to the sensorimotor level, allows a more helpful depiction of the content of the robot's experience. (Compare the "Anchored" c-knoxels in Chella, et al (2005, this volume)). But as it stands, the method may be too reductionistic on this second point. It seems desirable to have some way to distinguish notationally the experiences of a robot that produces those sensorimotor expectations directly, from that of a robot that has those same expectations as a result of an abstraction and de-abstraction process.

Stening, et al's "inversion" method allows them to compare the phenomenological world of a three-category robot to that of a five-category robot in a way that reveals the latter to be much more akin to how we experience the objective structure of space. But they also point out that the inversion method allows one to make relative comparisons that are essential for gauging the imaginative abilities of systems that have experiences that are fundamentally different from our own. Specifically, their method allows one to see that the three-concept robot's imagined world faithfully reconstructs its perceived world, even though both are radically different from how we would experience that space. A less subtle form of synthetic phenomenology, that merely focussed on the three-concept robot's inability to reconstruct *our* experience of the space, would have been unable to identify these successful aspects of the robot's imaginative capacities.

Synthetic phenomenology is also the focus of Gamez (2005, this volume). His review of the recent consciousness literature leads him to a position close to that of Prinz (2003) that "no test can separate out necessary and sufficient correlates or causes of consciousness. We can vary the ways in which the global functions of the brain are implemented in a vast number of ways, but since these will always lead to the same behavioural output… [and even to the same phenomenal experience from the first person perspective], any impact of these changes on consciousness cannot be measured and we will never know for certain whether a functionally… identical robot has conscious states at all." However, Gamez does not think this stance prohibits the development of a synthetic phenomenology. The paper develops an ordinal probability scale which is designed to be used in assessment of the possibility that our artificial creations might have consciousness. Gamez's contention is that the development of the field will eventually necessitate the research community and society at large to require just such a scale which will be of use in judging the development of the field both in its own terms, and for ethical purposes. For example, creating machines even with the strong likelihood of the capability of suffering might be intrinsically ethically problematic (but see section 9, below), and so it would be of great ethical import to be able to have some principled manner of assessment beyond personal intuition.

Having said that, Gamez's ordinal probability scale proposes formalizing our intuitions in a manner perhaps quite related to the axiomatic approach of Aleksander & Dunmall (2003). However, unlike those authors, Gamez is, as said before, skeptical about the possibility of developing strong axioms. Instead, building on Harnad's (1994) extension of the Turing test, Gamez proposes a metric for consciousness based on similarity to ourselves. The scale is thus strongly anthropocentric and by necessity will have difficulty accounting for other possible kinds of consciousness. Using the scale Gamez analyses several existing systems: Lucy (Grand, 2003); Demarse, Wagenaar, Blau, & Potter's Neurally Controlled Animat (2001); IDA (Franklin, Keleman, & McCauley, 1998); and, after Block (1978), a fictional functional system implemented by the population of China, in order to assess their respective likelihoods of being conscious.

8 Second-person approaches and neurophenomenology

The term 'neurophenomenology', (originating, like the 'enactive' approach, with Varela (1996; see also Thompson, Lutz and Cosmelli, 2005)), denotes the fusion of hermeneutic philosophy with rigorous empirical methods in neuroscience. A key element in neurophenomenology is the use of systematic techniques to enable phenomenologically trained subjects to give precise first-person accounts of features of their experiences. Second-person approaches, also favoured by Varela and others, stress empathetic interaction as a way of understanding consciousness. Social interaction – especially the notion that human consciousness develops from and is grounded in intersubjective processes – has been fundamental to the growth of first- and second-person studies in consciousness (Varela and Shear, 1999; Thompson, 2001). The sophisticated, interactive protocols being developed in Neurophenomenology may prove to be a source of data and design intuitions for the construction of systems that merit the attribution of phenomenological states. Since theorists are themselves social subjects, in giving an account of experience one cannot ignore the intersubjective relationships between theorist and subject (or robot).

The work of Nomura, Takaishi and Hashido (2005, this volume) has some relevance to this theme. They explore how virtual and robotic agents displaying many characteristics of consciousness (e.g. affective, empathic interactions) are perceived by participants in social settings such as psychotherapy and healthcare. The primary interest of Nomura and colleagues is in the psychological and sociological features of such applications. Their use of the term 'machine consciousness' stands somewhat in contrast to the more tentative use of those who regard machine consciousness somewhat as a 'holy grail' to be arrived at possibly only in the remote future. For Nomura et al., any system which is taken by (albeit naïve) users as possessing characteristics associated with conscious agents, may be taken to exemplify "machine consciousness" – so that even simple Eliza-style systems may display a schematic variety of that property. Even if "genuinely" or "literally" conscious machines lie in the realm of "science fiction" (as Shanahan would have it), the proliferation of computational agents displaying complex conscious-like characteristics that are taken by many to be signs of real consciousness may soon be a sociological fact. The social ramifications of the mass arrival of such pseudo-conscious agents are likely to extend over many other aspects of society than just therapeutic applications.

9 Ethical and legal issues

Some would argue that machine consciousness (unlike "mere" machine intelligence) has an inherently *ethical* dimension. A genuinely conscious machine (rather than one which merely shows outward signs or internal organizational features of consciousness) would perforce be capable of enjoyment, suffering, etc., and thus apparently be a genuine ethical subject (Torrance, 2000a; 2000b). If this is so, then the ethical dimensions of machine consciousness research can not really be treated as something external to the research enterprise. Rather, as we build increasingly complex artefacts in order to understand consciousness, normative concerns become essential, both to our understanding of the constitution of subjectivity, and to our appreciation of, and actions towards, the artefacts we create. These and other issues concerning the ethical import of machine consciousness are discussed by Torrance (2005, this volume).

Torrance cites the warning, expressed by Thomas Metzinger (2003), that, since being a conscious creature necessarily involves the possibility of great suffering, the development of artificially conscious creatures is perhaps an activity which we are morally obliged not to even start on. This may be a rather overzealous prohibition – our children will probably suffer to some degree or other during their lives, but we are surely not for that reason morally forbidden from procreating. But the point does lay down a strong challenge to strong machine consciousness researchers to become more aware of the ethical dimensions of their activities. The artificial consciousnesses we create won't be like our human children, and their differences from us may be profound and unpredictable.

Quite apart from the difficult moral and social questions raised by the machine consciousness enterprise there are also the legal questions. Calverley (2005, this volume) considers some of the relevant foundational issues in jurisprudence. He particularly considers the implications of debates between supporters of natural and positive conceptions of law, for the possible future emergence of artificial autonomous agents displaying features of consciousness. What extensions should be made within existing human-based legal – and moral – frameworks to properly take account of such agents? It seems clear that it will be necessary to clarify what kinds of legal *responsibilities* future autonomous machine consciousness agents might have, and also what legal *rights* we should accord them – what responsibilities we may have towards them. Calverley considers such questions in some depth, taking as his point of departure discussions that have already been initiated between cognitive scientists and lawmakers in the United States.

Acknowledgements

The authors would like to thank everyone who has assisted in the preparation of this work, including the E-Intentionality research group in COGS at the University of Sussex, the participants of the Skövde, Birmingham, Turin and Antwerp machine consciousness meetings, and the programme committee and authors for the AISB05 Symposium on Next Generation Approaches to Machine Consciousness.

References

Aleksander, I., & Dunmall, B. (2003). Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*, 10, 7-18.

Aleksander, I., Lahnstein, M., & Lee, R. (2005). Will and Emotions: A Machine Model that Shuns Illusions. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Block, N. (1978). Troubles with Functionalism. In C. Wade Savage (Ed.), *Minnesota Studies in the Philosophy of Science*, (Vol. IX). Minneapolis: University of Minnesota Press.

Bosse, T., Jonker, C. M., & Treur, J. (2005). Simulation and Representation of Body, emotion and Core Consciousness. In Chrisley, R., Clowes, R., and Torrance, S. (Eds.) *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Butterworth, G. (1998). A developmental-ecological perspective on Strawson's 'The Self'. In S. Gallagher & J. Shear (Eds.), *The Self.*

Calverley, D. J. (2005). Towards a Method for Determining the Legal Status of a Conscious Machine. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.* (pp. 75-84).

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

Chella, A., Frixione, M., & Gaglio, S. (2005). Planning by imagination in Cicerobot, a robot for museum tours. In Chrisley, R., Clowes, R., and Torrance, S. (Eds.) *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Chrisley, R. (1990). Cognitive map construction and use: A parallel distributed processing approach," in Touretzky, D., Elman, J., Hinton, G., and Sejnowski, T. (Eds.) *Connectionist Models: Proceedings of the 1990 Summer School.* San Mateo, CA: Morgan Kaufman. pp 287-302.

Chrisley, R. (1993). Connectionism, cognitive maps, and the development of objectivity. *Artificial Intelligence Review* **7**, pp 329-354.

Chrisley, R. (1995). Non-conceptual content and robotics: Taking embodiment seriously. In Ford, K., Glymour, C. and Hayes, P. (Eds.) *Android Epistemology.* Cambridge: AAAI/MIT Press, pp 141-166.

Clowes, R. W. (2003). Action Oriented Adaptive Language Games. Presented at the Third International Workshop on Epigenetic Robotics, Boston.

Cotterill, R.J. (1998). *Enchanted Looms: Conscious Networks in Brains and Computers.* Cambridge, UK: Cambridge University Press.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Papermac.

Damasio, A. R. (2000). *The Feeling of What Happens: Body, emotion and the making of consciousness*. Vintage.

Demarse, S., Wagenaar, A., Blau, A., & Potter, A. M. (2001). The neurally Controlled Animat: Biological Brains Acting with Simulated Bodies. *Autonomous Robots*, 11(3), 305-310.

Dennett, D. C. (1991). *Consciousness Explained.* Boston: Little, Brown..

Dennett, D. C. (2003). Who's On First? Heterophenomenology Explained. *Journal of Consciousness Studies*, 10, 19-30.

Franklin, S., Keleman, A., & McCauley, L. (1998). IDA: A Cognitive Agent Architecture. *IEEE*

International Conference on Systems, Man and Cybernetics, 3, 2646-2651.

Gallagher, S., & Meltzoff, A. (1996). The Earliest Sense of Self and Others: Merleau-Ponty and Recent Developmental Studies. *Philosophical Psychology*, 9, 213-236.

Gamez, D. (2005). An Ordinal Probability Scale for Synthetic Phenomenology. In Chrisley, R., Clowes, R., and Torrance, S. (Eds.) *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Gärdenfors, P. (2000). *Conceptual Spaces the Geometry of Thought*. London, England: MIT Press.

Gibson, J. J. (1962). Observations on active touch. *Psychological Review* (69), 477-491.

Grand, S. (2003). *Growing up with Lucy*. London: Weidendfield & Nicolson.

Haikonen, P. (2005). You Only Live Twice: Imagination in Conscious Machines. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Hanna, R. and Thompson, E. 2003. The mind-body-body problem. Theoria et Historia Scientiarum: *International Journal for Interdisciplinary Studies* 7.

Harnad, S. (1994). Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. In *Artificial Life 1* (Vol. 3).

Hesslow, G. (1994). Will neuroscience explain consciousness? *Journal of Theoretical Biology* **171**, 29-39.

Hurlburt, R. T. (1990). *Sampling Normal and Schizophrenic Inner Experience*. New York: Plenem Press.

Iizuka, H., & Ikegami, T. (2005). Emergence of Body Image and Dichotomy of Sensory and Motor Activity. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Ikegami, T. (2005). Consciousness, Emotion, and Imagination: A Brain-Inspired Architecture for Cogntive Robotics. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**, pp. 127-36.

Jonker, C. M., & Treur, J. (2002). Compositional Verification of Multi-Agent Systems: A Formal Analysis of Pro-Activeness and Reactiveness. *International Journal of Cooperative Information Systems*, 11, 51-92.

Libet, B., Gleason, C., Wright, E., & Pearl, D. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious imitation of free voluntary activity. *Brain*, 106, 623-642.

Maturana, H.R. and Varela, F.J. 1987. *The Tree of Knowledge. The Biological Roots of Human Understanding*. Boston: Shambala Press/New Science Library.

Metzinger, T. (2003) *Being No One: The Self-model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Noë, A. (2002). Is the visual world a Grand Illusion? *Journal of Consciousness Studies.* 9 (5/6), 1-12.

Nomura, T., Takaishi, K., & Hashido, T. (2005). Considerations of Machine Consciousness in the Context of Mental Therapy from Psychological and Sociological Perspectives. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

O'Regan, J.K. and Noë, A. (2001) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences,* 24 (5). 883-917

Picard, R. (1997). *Affective Computing*. Cambridge, Mass.: MIT Press.

Prinz, J. J. (2003). Level-Headed Mysterianism and Artificial Experience. In O. Holland (Ed.), *Machine Consciousness*. Exeter: Imprint Academic.

Ramachandran, V. S., & Blakeslee, S. (1998*). Phantoms in the brain*. New York: Harper Collins.

Searle (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences* **3**, pp. 417-24.

Shanahan, M. (2005). Consciousness, Emotion, and Imagination: A Brain-Inspired Architecture for Cogntive Robotics. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies* **10**:4-5, pp 133-172

Steels, L. (1998). Structural Coupling of Cognitive Memories Through Adaptive Language Games. In R. Pfeifer, B. Blumberg, J.-A. Meyer & S. Wilson (Eds*.), Animals to Animats 5: Proceedings of SAB 98* (pp. 263--269). Edinburgh: The MIT Press.

Stein, L. (1995). Imagination and situated cognition. In Ford, K., Glymour, C. and Hayes, P. (Eds.) *Android Epistemology*. Cambridge: AAAI/MIT Press.

Steels, L. (2003). Language Re-Entrance and the 'Inner Voice". In O. Holland (Ed*.), Machine Consciousness*. Exeter: Imprint.

Stening, J., Jacobsson, H., & Ziemke, T. (2005). Imagination and Abstraction of Sensorimotor Flow: Towards a Robot Model. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Stuart, S. (2005). The Binding Problem: Induction, Integration and Imagination. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Sugita, Y., & Tani, J. (2002). A connectionist model which unifies the behavioral and the linguistic processes. In M. I. Stamenov & V. Gallese (Eds.), *Mirror Neurons and the Evolution of the Brain (Vol. 42)*.

Thompson, E. ed. (2001). *Between ourselves. Second-person approaches to the study of consciousness.* Thorverton: Imprint Academic.

Thompson, E., Lutz, A. & Cosmelli, D. (2005) Neurophenomenology: An Introduction for Neurophilosophers in: A. Brook & K. Akins (Eds.) *Cognition and the Brain: The Philosophy and Neuroscience Movement.* Cambridge University Press.

Torrance, S (2000a) Ethics, Mind and Artifice, in R. Chrisley (Ed.) *Critical Concepts in Cognitive Science,* vol. 4. London: Routledge, 2000. (Reprinted from K.S.Gill (ed) *AI for Society.* Chichester: John Wiley, pp 55-72., 1986.)

Torrance, S. (2000b) Towards an Ethics for Epersons, in J. Barnden (Ed.) *Proceedings of AISB 2000 Symposium on Artificial Intelligence, Ethics and (Quasi-) Human Rights*, University of Birmingham, 47-52

Torrance, S. (2005). Thin Phenomenality and Machine Consciousness. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Trevarthen, C. (1994). The self born in intersubjectivity: The psychology of an infant communicating. In U. Neisser (Ed.), *The Perceived Self* (pp. 121-173.): Cambridge Univ. Press.

Varela, F. (1996) Neurophenomenology: A methodological remedy for the Hard Problem. *Journal of Consciousness Studies.* 3 (4). 330-349.

Varela, F. & Shear, J. (Eds.) (1999) *The View from Within: First-person approaches to the study of consciousness.* Thorverton: Imprint Academic.

Varela, F., Thompson, E. & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, 1991.

Vygotsky, L. S. (1986). *Thought and Language* (Seventh Printing edition). MIT Press.

Weber, A., & Varela, F. (2002) Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences,* **1**, pp. 97-125.

Wegner, D. W. (2002). *The Illusion of Conscious Will*. Cambridge MA: MIT Press.

Zlatev, J. (1999). The Epigenesis of Meaning in Human Beings, and Possibly in Robots*. Lund University Cognitive Studies*, 79.