International Journal of Machine Consciousness
Vol. 1, No. 1 (2009) 71–82

World Scientific
www.worldscientific.com

THE STRENGTH OF WEAK ARTIFICIAL CONSCIOUSNESS

ANIL SETH

Department of Informatics, University of Sussex,
Brighton, BN1 9QJ, UK
a.k.seth@sussex.ac.uk

Machine (artificial) consciousness can be interpreted in both strong and weak forms, as an instantiation or as a simulation. Here, I argue in favor of weak artificial consciousness, proposing that synthetic models of neural mechanisms potentially underlying consciousness can shed new light on how these mechanisms give rise to the phenomena they do. The approach I advocate involves using synthetic models to develop "explanatory correlates" that can causally account for deep, structural properties of conscious experience. In contrast, the project of strong artificial consciousness — while not impossible in principle — has yet to be credibly illustrated, and is in any case less likely to deliver advances in our understanding of the biological basis of consciousness. This is because of the inherent circularity involved in using models both as instantiations and as cognitive prostheses for exposing general principles, and because treating models as instantiations can indefinitely postpone comparisons with empirical data.

Keywords: Weak artificial consciousness; strong artificial consciousness; artificial life; explanatory correlate.

1. Introduction

The young field of machine consciousness (MC), or equally artificial consciousness (AC),[a] follows in the footsteps of artificial intelligence (AI) and latterly artificial life (AL). These sciences of the artificial take as their objective the attempt to recreate or represent in alternative media the basic biological phenomena of consciousness, intelligence and life, respectively. Two perspectives can be distinguished with regard to the scope and the feasibility of these projects.[1,2] On the "strong" view, the objective is to instantiate the phenomenon in question, in much the same way that artificial light is still light. There are as yet no credible examples of strong AC, nor any on the horizon. On the "weak" view, the objective is to develop synthetic simulation models of the necessary and perhaps sufficient mechanisms underlying the target phenomena in order to shed new light on the causal and explanatory links between these mechanisms and the phenomena they generate. The "weak" view is

[a]I prefer the term "artificial consciousness" over "machine consciousness" because it emphasizes the historical continuity with artificial intelligence and artificial life. They are otherwise entirely equivalent.

well illustrated by the use of computational models to understand the fluid dynamics underlying hurricane formation; the new understanding generated by these models carries no implication that hurricanes are in any sense computational.

The distinction between weak and strong artificial science has been hotly disputed in both AI and AL, and salutary lessons from these debates need to be well apprehended by MC/AC in order that this new field is not unfairly discounted in virtue of any residual skepticism about consciousness as a proper target for empirical science. This article sets out my own views on this issue, arguing for the scientific strength of weak AC in the context of the general utility of synthetic models in biological science. I also propose an agenda for weak AC which can be loosely described as "putting flesh on the bones of explanatory correlates of consciousness".

2. The Weakness of Strong Artificial Consciousness

2.1. *Is strong AC possible?*

Yes. Consciousness is a natural phenomenon. Its existence depends on the laws of physics, chemistry and biology, however imperfectly these are currently known. Accepting this plausible assertion implies that consciousness can *in principle* be generated by an artificially constructed device, operating in an appropriate context. But this is *not* to say that consciousness can necessarily be generated in or by computers, silicon devices, robots, or other objects constructed from arbitrary non-biological materials. For this to be true of necessity, "functionalism" must also be true. Functionalism is the theory that mental properties (including consciousness) are second-order properties constituted by their causal relations to one another and to sensory inputs and motor outputs.[3] For example, according to functionalism the mental state of being in pain is fully characterized by dispositions to say "ouch", to wonder whether one is unwell, to take an aspirin, and so on. However there is no *a priori* reason why functionalism need be true (or false), and therefore it is not yet clear what sort of artificially constructed device will be sufficient for giving rise to consciousness. Therefore, the plausible assertion that consciousness can be generated by an artificial device (let's say "machine") is rather trivial. It may in the end turn out that the only machines capable of giving rise to consciousness are indistinguishable from biological brains.

2.2. *Axiomatic approaches to strong AC*

In the absence of an accepted account of the sufficient neurobiological mechanisms underlying consciousness, most proponents of strong AC advocate an "axiomatic" approach in which criteria are established which, if fulfilled by an artificial device, warrant ascription of consciousness to that device. In a prominent example of this approach, Aleksander proposes a set of axioms based on introspectively derived features of consciousness, namely presence, imagination, attention, volition, and

emotion.[4] As Clowes and I have argued,[5] Aleksander's axioms can be challenged on at least two counts. First, axioms are defined in logic as propositions not proven but taken as self-evident. They express truths that can be taken for granted and which support prediction and explanation of other non-axiomatic phenomena. Aleksander's axioms, however, seem to represent *targets* for explanation rather than self-evident truths from which greater understanding flows. Second, the top-down organization of this axiomatic approach undermines claims of sufficiency, leading to a danger of trivial circularity. If a system is built to instantiate a set of axioms as stated, then it is said to *be* a conscious system. While the truth of such an assertion can be difficult to rule out *a priori* for any given axiomatic set, it then becomes difficult to know on what grounds to distinguish among competing sets of axioms.

A second example of the axiomatic approach is provided by the "information integration theory" (IIT) of Tononi.[6] According to Tononi, consciousness *is* information integration; a process or device is conscious to the extent that it is a single integrated entity with a (very) large repertoire of states, as measured by the information-theoretic quantity Φ (phi).[7] This statement, while perhaps not self-evident, is based on a carefully argued thought-experiment and therefore has at least the intention of being axiomatic. The essence of the thought-experiment is that conscious scenes are both unified (they are experienced "all of a piece") and differentiated (each scene is one among a large repertoire of possible scenes). In other words, conscious scenes are "complex" and each conscious scene constitutes a highly informative *discrimination* among many possible conscious scenes.[8–10] The coexistence of integration and differentiation (i.e., the complexity) underlying these discriminations is what is measured by Φ. Tononi contrasts conscious discriminations to those made by a simple photodetector (only two possible states, very little information) and by a digital camera (many possible states but no integration). The IIT offers an interesting contrast to Aleksander's approach. On one hand, the exclusive focus on information integration implies that many introspectively derived features of consciousness (sensory input and motor output, long-term or working memory, attention, self-reflection, language, emotion) are not in fact necessary for ascription of consciousness.[7] On the other hand, the use of information-theoretic concepts to bridge neural, mechanistic processes with conscious phenomenology allows that "information integration" can indeed help to *explain* features of consciousness, rather than presenting a further target for explanation.

Having said this, the IIT approach can itself be challenged. By identifying consciousness uniquely with information integration as measured by Φ, the theory implies that any system having sufficiently high Φ will be conscious, to that extent. This immediately raises the question: Can arbitrarily high Φ be obtained from trivial but sufficiently extensive systems? We have previously shown that simple fully connected Hopfield neural networks can in fact be arranged to generate arbitrarily high Φ when synaptic strengths are appropriately determined[11] (though see Ref. 12). The conclusion that even a Hopfield network could be arbitrarily (even infinitely)

conscious is naturally challenging. And even if one does accept this conclusion, it is not clear what extra understanding about human or animal consciousness is thereby gained.

It is perhaps more helpful for extending our understanding of consciousness if we consider information integration (or, in more neutral language, coexisting integration and differentiation, or complexity) as a deep structural and perhaps necessary property of any mechanism underlying consciousness, but not one that is by itself sufficient. Indeed, as I discuss further below (Sec. 3.2) there are alternative measures of complexity which have both practical and theoretical advantages over Φ as a measure of this sort of dynamical process.

2.3. *Pragmatic strong AC and lessons from artificial life*

So much for the "in principle" possibility of strong AC. What about the "in practice" scientific value of pursuing AC from this perspective? In considering this question, it is worth recalling the recent history of artificial life (AL), a research area in which processes and properties of living organisms are modeled or instantiated in alternative media. According to "strong" AL, simulations or models (computational, robotic, or otherwise) represent new *examples* of life: life "as it could be" in contrast to life "as it is".[13] According to "weak" AL, simulations/models are not considered to be alive themselves. Rather, they articulate precise hypotheses about the ability of particular causal mechanisms to give rise to target phenomena. At their simplest, models can be analytically solvable systems of equations, or thought-experiments, the consequences of which can be transparently determined. More often, models represent "opaque" thought-experiments in the sense that they instantiate mechanisms of sufficient complexity that their global causal effects cannot be determined except by explicit computational or numerical simulation.[14]

The debate between proponents of strong AL *versus* weak AL is now becoming rather sterile. It has been rendered increasingly moot firstly by a growing realization that "life" is a constellation concept; its ascription to a particular system depends to a greater or lesser extent on the presence of a range of properties including reproduction, metabolism, autopoeisis, autonomy, evolution, and the like.[15] In addition, the new field of "synthetic biology" is developing methods for artificially creating organisms using existing biological components[16]; this is genetic engineering taken to its logical extreme. The notion that such synthetic organisms are alive is less controversial than when applied to non-biological artifacts, simply because many features of the medium are preserved.

It may be that AC will follow a similar trajectory to AL with respect to the potency of the strong/weak distinction. Currently, however, concepts of consciousness are still insufficiently elaborated as compared to concepts of life, though consciousness may indeed have a constellation nature.[17] Moreover, "neurobiological engineering" is at an embryonic stage and examples of potentially conscious "synthetic neural systems" are

poorly developed as compared to examples and methods of synthetic biology and genetic engineering. (Simple synthetic neural systems have been created (see Ref. 18), however, no claims are made that such systems are in any sense conscious.)

Perhaps more important than the ontological status of AL models are the epistemological uses to which they are put. If AL models are accepted, provisionally or otherwise, as (strong) examples of life, rather than as (weak) simulations of living systems, then the data acquired from these models has the same empirical status as data acquired from actual, biological, living systems. One advantage of adopting this stance is that models are almost always easier to understand than their biological counterparts.[19] They are generally composed of fewer components, which interact according to known rules; models can be run repeatedly with the same or different starting conditions. Most importantly, all parts of a model system are available to observation and manipulation. However, these features are benefits of all types of models in all situations.[20] Therefore, the only additional advantage in treating AL models as examples of living systems is that new insights about life can apparently be extracted *directly* from the model, rather than indirectly by comparison of model behavior with empirical data from real systems. But the value of doing this in turn rests on one's conviction that the model system is (ontologically) alive, and there is an inherent circularity in using models to expose principles of life that can serve as criteria for separating living from non-living systems, when one's doing so depends on accepting *a priori* that one's model is an example of the former.[21]

A second and more controversial advantage is that AL models can *depart* from their biological counterparts in all sorts of ways. This property is reflected in the AL slogan "life as it could be".[13] The intuition is that existing, biological, examples of life consist of both fundamental features, essential to any living system, and contingent features, that have been acquired during evolution in response to diverse selective pressures and which are not strictly necessary for life *per se*. Because strong AL models are not necessarily saddled with contingent features, the argument goes that examination of their behavior can help identify "general principles" underpinning living systems of all kinds. (More than 30 years ago, Dennett proposed to "*make up* a whole cognitive creature" in order to uncover "*very* general, *very* abstract principles".[22]) However, for such general principles to extend our understanding of biological life, it remains an ultimate requirement that comparisons with empirical data are made.[19] Unfortunately, by taking an AL model as a realization of a living system, such comparisons can be deferred indefinitely. Moreover, as soon as models depart by design from their biological counterparts, criticisms of any particular model element as arbitrary or unrealistic can be more readily deflected.

The lesson for AC is straightforward. Setting to one side the question of whether models of neural (or other) mechanisms are actually (ontologically) conscious, we should be wary of even provisionally treating such models as if they instantiate consciousness. Such an epistemological stance can lead to a trivial circularity of analysis and understanding, in which models are treated as both realizations and as

cognitive prostheses, and to a perpetual deferment of ultimately necessary comparisons of model behavior with empirical data from biological conscious systems.

3. The Strength of Weak Artificial Consciousness

3.1. *Weak AC and synthetic modeling*

I turn now to positive arguments in favor of weak AC. At bottom, these arguments are no different from those offered in favor of synthetic modeling in almost any domain: Synthetic models can articulate specific hypotheses through carefully constrained simulation and in doing so they allow otherwise impenetrable phenomena to be elucidated. The implication is that weak AC is not fundamentally different from other more mainstream forms of computational modeling in biology, psychology, and neuroscience.

Valentino Braitenberg, much revered in theoretical biology and neurobiology, proposed with his law of "uphill analysis versus downhill synthesis" that complex phenomena that resist direct analysis can best be understood by analysis of less complex alternatives instantiated in simulation.[23] As already mentioned, one reason this is true is that models are easier to observe and manipulate in arbitrary ways and with arbitrary detail and precision (they are also, in general, less complex). However, because on the weak approach, simulations are treated *as* simulations, and not as instantiations or realizations, a key criterion for the ability of synthetic models to contribute to scientific understanding is that their mechanisms and behavior must, on some level, be comparable to the target biological phenomenon.[19] Without such comparisons, models risk floating free from their empirical anchors and being evaluated instead on less reliable criteria such as analytical tractability, parsimony, elegance, etc.

Useful comparisons to empirical data can take place at many different levels of description. For example, if gamma-band synchrony is considered to be important for consciousness,[24] models can be constructed that exhibit such synchrony and then other of their properties can be observed and compared with empirical data.[25] Alternatively, models can be constructed to exhibit high-level, functional properties associated with consciousness, such as rapid discrimination among multiple scenes[9] (Tononi and Edelman's "dynamic core"), or flexible integration of and selection among multimodal sensorimotor signals[26] (Baars' "global workspace" theory). Correspondence of such high-level properties with those observed empirically then warrants further comparisons at the level of mechanism. For example, Dehaene and colleagues have proposed a model in which sensory stimuli mobilize excitatory "workspace" neurons with long-range corticocortical axons, leading to the emergence of global activity patterns among these neurons.[27] This model is given as a neural implementation of global workspace theory and its analysis leads to various experimental predictions, for example, that consciousness of a stimulus is a non-linear ("all-or-nothing") function of stimulus salience.

3.2. *From properties to criteria*

A good scientific theory requires both *criteria* for deciding the admissibility and relevance of data, or to guide the construction of simulations and/or artifacts, as well as clearly defined *properties* (explananda) to which this data should relate.[5,28] The difference between a criterion and a property is one of testability and/or implementability; a criterion is *operational*. A testable property can naturally be treated as a criterion. For example, consciousness has the property of irregular, low amplitude EEG activity.[28] Being testable, this property can serve as a criterion for evaluating empirical evidence for consciousness. Consciousness also has the property of subjectivity; however it is not clear that subjectivity is something that is either present, or not present, in empirical data. Note that properties, as described here, are distinct from "axioms" (see Sec. 2.2) because properties can be revised as new theoretical insights and new experimental data become available. For example, the concept of subjectivity may change as we understand more about how conscious states are modulated by the complex interplay of egocentric and allocentric representations across brains, bodies, and environments.[29]

Clowes and I have argued that weak AC models can contribute to consciousness science specifically by transforming properties into testable, operational criteria.[5] A good example is provided by Dehaene's neuronal model of global workspace theory already mentioned. Another example is provided by the property of combined integration and differentiation (complexity) within conscious scenes.[9] A series of measures have now been proposed which operationalize this property such that it can be tested for empirical data. These measures include "neural complexity"[9] and "causal density"[30] (Φ fits uneasily in this category because it is not feasible to measure for non-trivial systems, and because it is explicitly asserted as sufficient for consciousness) (see Refs. 6, 11, 31 for reviews). Weak AC approaches could indeed go further, by developing explicit computational models which are analyzed in terms of neural complexity and causal density, supporting comparisons with empirical data at multiple levels of description.

A more subtle role for weak AC is to show that apparently distinct properties of consciousness may arise from common neural mechanisms.[5] For example, consciousness has the properties of "global availability" (the same content is available to different neurocognitive processes) and "adaptivity" (conscious contents are always oriented toward flexible behavioral control). Neuronal global workspace models show how these two properties can arise from a common mechanism, namely the ongoing competition for scarce resources within a globally distributed workspace. In another example, future mechanistic models of neural complexity and causal density may connect the property of compexity with that of metastability, i.e., the property that each conscious scene shades naturally, and over a predictable timescale, into another related yet distinct conscious scene.[5] In this way, weak AC encourages the malleability of our conception of relevant properties of consciousness in a manner not possible within axiomatic approaches.

3.3. *Structural properties and explanatory correlates of consciousness*

What properties of consciousness should weak AC models target? Consciousness science may be best served by focusing on "structural" properties of consciousness.[5,10,11] These are aspects or dimensions of the way the world is presented to us through conscious experience, as opposed to particular examples of canonical conscious experiences (e.g., the experience of redness). Structural properties of consciousness include the coexistence of integration and differentiation (complexity),[9] the reference of conscious contents to a subjective first-person perspective,[32] the shaping of conscious states by emotional and mood states,[33] and the association of consciousness with intention, agency, and volition.[34] (See Refs. 8, 28, 32 for more comprehensive lists.) Accounting for these structural properties in terms of neural mechanisms would represent a major achievement for consciousness science, and initial steps have already been taken with the proposal of operational measures of complexity.

The elaboration of neural mechanisms underlying structural properties can be characterized as the search for *explanatory correlates of consciousness* (ECCs).[10] This concept is a useful extension of the notion of a "neural correlate of consciousness", which refers to activity within brain regions or groups of neurons having privileged status in the generation of conscious experience.[35,36] The development and identification of ECCs represents an important transition from correlation to explanation. In other words, ECCs attempt to explain *why* particular neural correlates have the privileged relation with phenomenal experience that they do.

Weak AC models are well placed to develop ECCs by exposing the global dynamical properties of particular neural mechanisms. Although ECCs relating to complexity are perhaps the most advanced, others are now beginning to take shape, catalyzed in each case by weak AC models. For example, Grush has described a framework based on forward modeling which hints at the mechanistic origins of the first-person perspective.[37] This framework is based on learning forward-models of body-environment interactions; the models are driven by efference copies of motor commands and provide expectations of sensory feedback. Other related models have been developed by Bongard[38] and Holland,[39] each of which involve the explicit development of internal self-models of body-environment relations. Weak AC models directed at emotional or volitional experience are much scarcer, though tentative first steps have been taken (see Ref. 10 for a review).

4. Discussion

Weak AC consists of two mutually reinforcing activities. First, the construction and analysis of synthetic models (computational and/or robotic) can help connect neural dynamics to structural properties of conscious experience. By serving as a bridge between properties and potential mechanisms, weak AC models can (i) transform structural properties into operational criteria, and (ii) unify structural properties by

showing their dependence on a common set of underlying mechanisms. Second, theoretical approaches can help define ECCs, whose properties and experimental predictions can be explored through the subsequent construction of synthetic models. An important criterion for successful weak AC models is that their outputs must allow comparison either with putative biological mechanisms underlying consciousness, or with structural phenomenological properties, or in the best case with both simultaneously.

Strong AC, while superficially more ambitious than weak AC, may in practice result in less or no progress in scientific understanding. One reason for this is the circularity involved in treating a model as an instantiation which can be analyzed to reveal general principles, in the absence of knowledge of just these general principles that would validate the interpretation of the model as an instantiation in the first place. Another is that by building strong AC models which deliberately depart from their biological counterparts, comparisons with empirical data can be perpetually postponed and criticisms of any particular element of a model can be all too easily deflected.[19]

Other varieties of AC that can be distinguished lie between the extremes of "weak" and "strong". For example, Chrisley's notion of "lagom" AC involves a "necessary but not sufficient" attitude towards AC.[40] At least two kinds of lagom AC are possible. Chrisley's point in introducing lagom AC was to highlight a kind of model that captures the *formal structure* of the causal interactions underlying consciousness, without these models being assumed to thereby instantiate consciousness. This reading is only distinct from weak AC (as described here) if one assumes that weak AC is concerned only with replicating conscious behavior, and not with explaining the actual causal, mechanistic properties that underlie consciousness. However, the weak AC I have argued for here is in fact explicitly aimed towards exposing explanatory connections between causal, mechanistic factors and functional, behavioral, and phenomenal properties, and is therefore already well aligned with Chrisley's lagom AC.

It follows that there is another kind of lagom AC, distinct from Chrisley's original focus, which consists in a necessary-but-not-sufficient attitude towards the *modeling medium*. According to this new version of lagom AC, there is non-arbitrariness with respect to the modeling medium, not only with regard to instantiating consciousness, but also for explaining how consciousness arises in nature. For example, it may turn out that exposing causal relations between neural mechanisms and structural properties of consciousness will require a synthetic model that is itself composed of neurons or some other non-arbitrary material. (The recent emphasis on fully (physically) embodied neurorobotic models can be seen as an example of this intuition, albeit with respect to the importance of physical bodies rather than neurons.[39]) And yet it may simultaneously be the case that such a medium-specific model, although thereby explanatory, nonetheless fails to instantiate consciousness itself, hence the term "lagom".

5. Conclusions

The AC project may appear to be doubly hamstrung. Not only is the concept of "artificial X" controversial for many instances of X (including, prominently, life and intelligence), but the scientific study of consciousness itself is only now re-emerging as a legitimate enterprise. Here, I have attempted to chart a course for AC that avoids these potential pitfalls. I have argued that weak AC represents a sensible application of generally accepted synthetic modeling practices and should be distinguished from, and preferred to, strong AC. I have also shown how the concept of an "explanatory correlate" provides a mechanism by which synthetic models of neural processes can account for — rather than merely correlate with — structural properties of conscious phenomenology.

I would like to end with a speculation: A successful weak AC programme may, in the end, deliver much of what strong AC promises. As one progressively builds in new constraints that counter theoretical objections or accommodate empirical mismatches that arise from existing models, so the models in question may actually tend towards the instantiation of systems that might genuinely be considered conscious.[5] If it is possible to establish the mechanistic basis for every generally accepted structural property of consciousness, then it becomes plausible that there will be no further limitations on building an "actually conscious" entity. Of course, as noted at the outset, and as suggested by the useful notion of lagom AC, it may well turn out that an AC model that is sufficiently rich to account for all structural properties of consciousness will not be implementable in computers or robots, and will instead require implementation in neural or some other material.

Acknowledgments

I am grateful to Prof. Antonio Chella for inviting this contribution. Its composition was supported by EPSRC Leadership Fellowship EP/G007543/1. Thanks are also due to Ron Chrisley for clarifying aspects of lagom AC for me.

References

1. O. Holland, *Machine Consciousness* (Imprint Academic, 2003).
2. J. Searle, Minds, brains, and programs, *Behavioral and Brain Sciences* **3** (1980) 417−457.
3. J. A. Fodor, *Psychosemantics* (MIT Press, 1987).
4. I. Aleksander and H. Morton, Depictive architectures for synthetic phenomenology, in *Artificial Consciousness*, eds. A. Chella and R. Manzotti (Imprint Academic, 2007), pp. 67−81.
5. R. W. Clowes and A. K. Seth, Axioms, properties and criteria: Roles for synthesis in the science of consciousness, *Artif. Intell. Med.* **44** (2008) 91−105.
6. G. Tononi, An information integration theory of consciousness, *BMC Neurosci.* **5** (2004) 42.
7. C. Koch and G. Tononi, Can machines be conscious? *IEEE Spectrum*, 2008.
8. G. M. Edelman, Naturalizing consciousness: A theoretical framework, *Proc. Natl. Acad. Sci. USA* **100** (2003) 5520−5524.
9. G. Tononi and G. M. Edelman, Consciousness and complexity, *Science* **282** (1998) 1846−1851.

10. A. K. Seth, Explanatory correlates of consciousness: Theoretical and computational challenges, *Cognitive Computation*, **1** (2009) 50–63.
11. A. K. Seth *et al.*, Theories and measures of consciousness: An extended framework, *Proc. Natl. Acad. Sci. USA* **103** (2006) 10799–10804.
12. D. Balduzzi and G. Tononi, Integrated information in discrete dynamical systems: Motivation and theoretical framework, *PLoS. Comput. Biol.* **4** (2008) e1000091.
13. C. Langton, Artificial life, in *Proc. Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*, Vol. 6, 1989, pp. 1–48.
14. E. Di Paolo *et al.*, Simulation models as opaque thought experiments, in *Artificial Life VII: 7th Int. Conf. Simulation and Synthesis of Living Systems*, 2000.
15. M. Boden, ed. *The Philosophy of Artificial Life* (Oxford University Press, 1996).
16. D. G. Gibson *et al.*, Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome, *Science* **319** (2008) 1215–1220.
17. A. Zeman, What in the world is consciousness, *Progress in Brain Research* **150** (2005) 1–10.
18. T. B. Demarse *et al.*, The neurally controlled animat: Biological brains acting with simulated bodies, *Auton. Robots* **11** (2001) 305–310.
19. B. Webb, Animals versus animats: Or why not model the real iguana? *Adaptive Behavior.* (in publication).
20. J. L. Krichmar and G. M. Edelman, Brain-based devices for the study of nervous systems and the development of intelligent machines, *Artif. Life* **11** (2005) 63–77.
21. B. L. Keeley, Shocking lessons from electric fish: The theory and practice of multiple realization, *Philosophy of Science* **67** (2000) 444–465.
22. D. Dennett, *Brainstorms* (MIT Press, 1978).
23. V. Braitenberg, *Vehicles: Experiments in Synthetic Psychology* (MIT Press, 1984).
24. A. K. Engel and W. Singer, Temporal binding and the neural correlates of sensory awareness, *Trends Cogn. Sci.* **5** (2001) 16–25.
25. E. M. Izhikevich and G. M. Edelman, Large-scale model of mammalian thalamocortical systems, *Proc. Natl. Acad. Sci.* **105** (2008) 3593–3598.
26. B. J. Baars, *A Cognitive Theory of Consciousness* (Cambridge University Press, 1988).
27. S. Dehaene *et al.*, A neuronal network model linking subjective reports and objective physiological data during conscious perception, *Proc. Natl. Acad. Sci.* **100** (2008) 8520–8525.
28. A. K. Seth *et al.*, Criteria for consciousness in humans and other mammals, *Consciousness and Cognition* **14** (2005) 119–139.
29. B. Lenggenhager *et al.*, Video ergo sum: Manipulating bodily self-consciousness, *Science* **317** (2007) 1096–1099.
30. A. K. Seth, Causal connectivity analysis of evolved neural networks during behavior, *Network: Computation in Neural Systems* **16** (2005) 35–55.
31. A. K. Seth *et al.*, Measuring consciousness: Relating behavioral and neurophysiological approaches, *Trends Cogn. Sci.* **12** (2008) 314–321.
32. T. Metzinger, *Being No One* (MIT Press, 2003).
33. A. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harvest Books, 2000).
34. P. Haggard, Human volition: Towards a neuroscience of will, *Nat. Rev. Neurosci.* **9** (2008) 934–946.
35. G. Rees *et al.*, Neural correlates of consciousness in humans, *Nat. Rev. Neurosci.* **3** (2002) 261–270.
36. G. Tononi and C. Koch, The neural correlates of consciousness: An update, *Ann. NY Acad. Sci.* **1124** (2008) 239–261.

37. R. Grush, The emulation theory of representation: Motor control, imagery, and perception, *Behav. Brain Sci.* **27** (2004) 377–396.
38. J. Bongard *et al.*, Resilient machines through continuous self-modeling, *Science* **314** (2006) 1118–1121.
39. O. Holland, A strongly embodied approach to machine consciousness, *J. Consci. Stud.* **14** (2007) 97–110.
40. R. Chrisley, Philosophical foundations of artificial consciousness, *Artif. Intell. Med.* **44** (2008) 119–137.