ψυχη
PSYCHE

Review of P. O. Haikonen, *The Cognitive Approach to Conscious Machines*

Mitch Parsell
Department of Philosophy
Macquarie University
Sydney, NSW 2109

mparsell@scmp.mq.edu.au

PSYCHE 11 (2), March 2005

Review of: Pentti O. Haikonen (2003). *The Cognitive Approach to Conscious Machines*. UK: Imprint: ISBN 0907845428.

Haikonen (2003) is an attempt to explicate a platform for modelling consciousness. The book sets out the foundational concepts behind Haikonen's work in the area and proposes a particular modelling environment. This is developed in three parts: part 1 offers a brief analysis of the state of play in cognitive modelling; part 2 an extended treatment of the phenomena to be explained; part 3 promises a synthesis of the two preceding discussions to provide the necessary background and detail for the proposed modelling environment. This final part covers a broad range of technical details from the nature of the representational-computational economy instantiated, to the control of motor output, to the means of implementing emotions in artefacts. Haikonen proposes an environment based on a distributed representational economy, instantiated in a neural network architecture and trained using associative learning regimes, but which also has symbolic processing abilities to handle the critical task of generating inner language.

Defending a modelling environment requires a demonstration that one's preferred platform has sufficient resources model the target phenomena, in this case consciousness. Haikonen approaches this from a particular pragmatic perspective. He is not concerned with philosophical complexity or even theoretical questions *per se*, but with practical questions concerning modelling and implementation. I term this an "engineering approach." The attention to practical issues sees claims advanced as modelling *hypotheses*, rather than conclusions to persuasive lines of argument. Indeed, Haikonen (p. X) is dismissive of theoretical/philosophical justification:

> Philosophers have the luxury of presenting their work as theoretical questions. … We engineers do not have this luxury. Our ideas that first appear as a design philosophy must face the acid test of practicality.

Though a refreshing perspective, his engineering approach is under-developed. Indeed, it face two difficulties that compound each other: (i) the engineering approach itself is prone to neglect the justificatory grounding of hypotheses and (ii) Haikonen is a little hasty with detail. I will return to these global criticisms. First, I wish to draw out the major conclusions of each part.

Part 1 (pp. 9-37) provides an engaging and non-technical summary of the cognitive modelling literature. Good Old-Fashioned Artificial Intelligence (GOFAI) is presented with an historical eye, moving from 1940s information "Zeppelins", through Turing and von Neumann, to artificial minds and objections to AI. Connectionist models are introduced as abstract and simplified biological networks. Connectionism is Haikonen's preferred paradigm, though he claims the ability to process information symbolically is necessary to consciousness: his model depends on the inner *re*-presentation of linguistic content, a task assumed to require symbolic processing. Thus he proposes a connectionist architecture which realises other computational capacities (pp. 36-37). In itself, this is a reasonable proposal, though the case for needing symbolic processing is under-argued. Further, though the discussion in this part is informative, it is rarely disinterested. The discussion of the problems with traditional AI is much better than the discussion of problems with connectionism. The problems facing connectionism are either glossed over—in the case of the AI founded attack (eg, Fodor and Pylyshyn, 1988; Fodor and McLaughlin, 1990)—or completely ignored—in the case of the more radical extended, embedded and embodied mind based attack (eg, Brooks, 1991b, Clark and Chalmers, 1998). Indeed, connectionist nets are presented as the solution to the problems plaguing traditional AI (chapter 2, "Artificial Neural Networks to the Rescue"). Despite this rhetorical spin, all but one of the problems with AI are equally problems for connectionism. Further, the criticisms of both GOFAI and connectionism emerging from the radical dynamic and extended platform (eg, van Gelder, 1999 and 1996; Garson, 1996; Foss 1992) (eg, Brooks, 2002, 1997, 1991a and 1991b)—criticisms to which Haikonen's own proposal is subject—are not treated.

Part 2, which occupies five chapters (pp. 41-162), introduces the broad range of phenomena that fall under the consciousness umbrella. Haikonen moves from general considerations about how perception, cognition and consciousness should be studied scientifically, to philosophical accounts of the nature of our mental life, through consideration of possible learning regimes, to finish with a direct consideration of conscious experience. Although certain positive claims are under-argued (I will note some below), I find this part of the book the most satisfying. The most significant and central conclusion concerns the use of language. Haikonen regards language as not merely a communicative device, but also as a tool for thought. The defence of this begins with a consideration of two traditional linguistic theories of mind, namely Fodor's (1975) language of thought (LOT) hypothesis and natural language theories (pp. 127-131). LOT is dismissed for its failure to deliver a plausible account of acquired representational capacities and for explanatory poverty (i.e., it merely displaces rather than answers critical questions). I think Haikonen is right to point to both errors, but his treatment of natural language theories is less compelling. Despite this, Haikonen's positive account is interesting. It is centred on a multimodal model of language in which each perceptual modality has a separate representational space, but with representations in distinct modalities can be connected to each other (p. 131). Memories are associatively encoded such that any given representation can evoke other related representations (p. 133).

In part 3, Haikonen attempts to pull together the threads of the preceding two parts. The discussion begins with an examination of the modelling abilities of artificial neural networks, which is followed by chapters expanding this general proposal to cover models of perception (chapter 12), motor function (chapter 14), emotions (chapter 15), language and inner speech (chapter 16) and imagery and thinking (chapter 17). The penultimate chapter (18, pp.247-261) deals with machine consciousness specifically and considers the application of Haikonen's model to the phenomenon of blind-sight (p.254). The best treatment in this section, and indeed the book, is the discussion of emotions (chapter 15; pp. 211-218). In Chapter 6 (pp. 99-123), Haikonen defends a non-representational systems reaction account of pleasure and pain, in which both elicit attention, but in substantially different ways. Pain disrupts attention within perceptual modalities, while pleasure aims at sustaining attention via the relaxation of un-related modalities (pp.101-105). Emotional states are supposed to result from combinations of multiple simultaneously occurring system reactions. For example, *horror* is the combination of the systems reactions bad, novelty, and withdrawal; *curiosity*, novelty and approach; and, *fear*, pain, bad and withdrawal (see figure 6.4, p. 114; and table 6.4, p. 115 for greater detail and more examples, respectively). The proposed artificial system instantiates machine emotions as combinations of primitive system reactions, which occupy the same functional role in the machine as basic reactions in biological organisms, together with the cognitive evaluation of those reactions and their perceived causes. This is a clearly stated hypothesis; moreover, the engineering detail is spot on. There is, however, a methodological complexity here. Suppose the basics of the model are correct: How should we individuate, identify and measure the basic systems reactions? Similarly, how should we establish which combinations produce which emotions? Armchair answers seem ruled out. Hence, these questions must be open to empirical study. If so, Haikonen owes us at least a first pass at how such study will occur. This is not provided. This failure is indicative of the problem with the engineering approach generally.

As noted, the engineering perspective results in proposals being set forth as modelling hypotheses. In many circumstances, hypotheses are fine. One needs no argument in support of a hypothesis one is setting forth to test, for the test itself is the "argument". In such situations, all that is required is a clear, testable statement of the hypothesis. But when the hypotheses in question are global claims about the correct platform for modelling consciousness more is needed -- ideally, a demonstration that the proposed environment is the only possible or the best available platform. Of course, such arguments are difficult to come by, but minimally, there needs to be a demonstration that the proposed platform is (i) plausible and (ii) at least as good as the alternatives. Haikonen never gives anything approaching the ideal, and, indeed, rarely moves beyond (i) to (ii). To be sure, some of the hypothesis could turn out to be true, but Haikonen neither demonstrates their truth nor justifies their *relative* plausibility. Consider the discussion of inner speech. Haikonen never successfully dismisses the natural language view, there is no argument to support the multi-modal view over the natural language view and nothing to show that either entails a symbolic representational economy. In some areas, Haikonen never adequately states his hypotheses, so he fails to achieve (i). For example, Haikonen equivocates between semi- or a fully-distributed (superpositional) data-encoding systems. Within the chapter devoted to the representation of information (10, pp. 169-173) it appears that Haikonen is arguing for a fully-distributed system. But in later chapters he seems committed to representational atoms. This is incompatible with the deeply context sensitive nature of fully-distributed representation (see Clark, 1989). For example, in chapter 14 (pp. 203-209) on motor function, distributed representation is supposed "suitable for motor *primitives*" (p. 203; emphasis added), where the primitives are hard-wired responses (p. 205). Moreover, Haikonen is committed to the necessity of symbolic processing. The principle of charity favours a semi-distributed interpretation of Haikonen's position, Haikonen's appeal to the advantages of distributed encoding may not be available to semi-distributed models.

While not fully persuasive, Haikonen's most interesting and radical proposals concern the modelling of a perspective and time perception. In chapter 4, on cognition and perception, Haikonen proposes a model for the development of a point of view. After an illuminating and extended discussion of the "location" of sensations, we are introduced to the "headphone illusion" in which in-the-ear phones create the illusion that the sound is eminating from within one's head. This and other "dis-locations" (this is not his term) of sensations are supposed to form the basis of our ability to occupy a perspective. The perceived location, together with our knowledge of the actual location—or strictly the location of the cause—enables a dis-entangling of our perceptual organs and, by extension ourselves, from the world. This in turn creates a point of view: "…the vantage point of view arises from the ability to attribute to sensations a point of origination that is different from the actual sensory nerve end position …" (p. 74). An intriguing proposal, but I am not convinced. Indeed, I find it hard to see how any form of *knowledge* could possibly do the necessary work. Haikonen is aiming for an explanation of why it is that we always occupy a particular perspective. It does not seem to me plausible that knowledge of the location of causes could ground such a point of view. Empirical knowledge of this form seems too far removed—at too high a level—to affect something so fundamental as our position in the world. Haikonen provides no argument to convince otherwise.

Chapter 5 (pp. 75-98), concerning learning, reasoning and memory, presents a model of the perception of the flow of time. We perceive time as moving forward into the future "because today we have more memories than yesterday and we know that today's memories are more recent than yesterdays" (p. 85). Perhaps this claim could be plausibly supported, but Haikonen makes no such attempt and background issues—such as the nature of memory, for example—are dealt with at nothing more than the most trivial level. We are told, for example, that the vividness of memories may help explain their perceived temporal ordering, but that this need not reflect the objective order at which they occurred. These claims seem at odds and nowhere is reconciliation attempted. There is a neurological mechanism proposed to explain time perception, but it comes in the form of a metaphor. The metaphor trades on "recording": the more novelty in a situation, the more detail recorded, hence, the bigger the data-slice associated with the period, thus the longer the perceived time. What is needed here is a rigorous/systematic account of novelty and a metric for data encoding, but it is doubtful either will arrive soon. There is, of course, one big benefit of this account of time perception for an engineer: a machine could be given such a sense should it have the ability to create and store its own personal history.

In the final chapter of part 2 (chapter 8; pp. 141-162) we confront consciousness itself, beginning with a brief tour of the main philosophical theories of consciousness (or strictly, the mind-body relation): dualism, materialism and idealism. We are introduced to Chalmers' easy and hard problem division of the domain. Haikonen argues that the easy problem of qualia—explaining their intentional character—is solvable via a causal theory of content. For example, certain signals depict redness because of their causal origins. The well-know problem of mis-representation—the inability *in principle* to explain how a representational system can fail, given it is supposed to represent whatever causes tokenings to arise—that is fatal to such a straightforward causal account of meaning is left untreated (see Fodor, 1990). The hard problem—explaining the "feel" of qualia—is supposed solvable by the previously proposed model of emotions. At this point Haikonen considers a possible counter argument. Suppose that it is true that qualia are linked to systems reactions. Even if we accept this, our interlocutor continues, why maintain that they "feel like something" to the system itself? Haikonen's response is to suggest that this is an empirical research question: 'to know for sure elaborate investigation would be needed' (p. 148). We are told that critical study would involve experimentation on an artificially conscious machine, but the nature of such an experiment remains unexplained and without such detail, I don't see how the "experiment" would be anything but an extension of the Turing test. The chapter finishes with a test for theories of conscious that parallels Johnson infamous "refutation" of Berkeley's idealism, but instead of toe kicking it involves hammer hitting. Here is the test: hit your thumb with a hammer and now think about theories of consciousness. The resulting experience is supposed to refute the linguistic theory of consciousness, social accounts of consciousness and theories requiring memory. I will let you make what you will of this test.

Overall, Haikonen provides both a promising approach and a plausible modelling proposal. There are, nevertheless, some serious problems with the text. The project and main conclusions, though intriguing and worth some serious thought, are inadequately defended. As an engineering proposal, the book would seem to lack the punch to persuade other researchers of the prospects of the proposed platform. But so long as one is after intuition pumps and hypotheses, rather arguments and well-supported conclusions this book should satisfy. As such, it is recommended for serious researchers in modelling consciousness. This book would serve as a useful introduction to the cognitive modelling of consciousness for those more familiar with cognitive modelling than with consciousness, but because of the skewered presentation in part 1 I would not recommend it for those more familiar with the study of consciousness than cognitive modelling.

Acknowledgements

Thanks to Tim Bayne and Cynthia Townley for very helpful comments on an earlier version of this review, Jay Garfield, Mark Colyvan and John Sutton for many discussions on related issues and Nik Fish for inspiration.

References

Brooks, R. A. 1991a. New Approaches to Robotics. *Science 253*: 1227–1232.

Brooks, R. A. 1991b. Intelligence Without Representation. *Artificial Intelligence Journal 47*: 139–159.

Brooks, R. A. 2002. *Flesh and Machines*, Pantheon Books, New York (NY).

Brooks, R.A. 1997. From Earwigs to Humans. *Robotics and Autonomous Systems 20* (2–4): 291–304.

Clark, A. & Chalmers, D. 1998. The Extended Mind. *Analysis 58*: 7-19

Clark, A. 1989. *Microcognition*. MIT Press: Cambridge (MA).

Fodor, J. & Pylyshyn, Z. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition 28*, 3-71.

Fodor, J. &McLaughlin, B. 1990. Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work. *Cognition 35*: 183-204

Fodor, J. A. 1975. *The Language of Thought*. Harvard University Press : Cambridge (MA).

Fodor, J. A. 1990. *A Theory of Content and Other Essays*. MIT Press Cambridge (MA).

Foss, J. 1992. Introduction to the Epistemology of the Brain: indeterminacy, micro-specificity, chaos, and openness. *Topoi 11*: 45-57

Garson, J. 1996. Cognition Poised at the Edge of Chaos: A Complex Alternative to a Symbolic Mind. *Philosophical Psychology 9*: 301-322

van Gelder, T. 1995. What Might Cognition be, if not Computation? *Journal of Philosophy 91*: 345-381

van Gelder, T. 1999. Dynamic Approaches to Cognition. In R. Wilson & F. Keil (eds). *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press: Cambridge (MA): 244-246