[54] H. A. Simon. *Sciences of the Artificial.* M.I.T. Press, Cambridge MA, 1969.

[55] A. Sloman. Beyond turing equivalence. In *Proceedings Turing 1990 Colloquium*, 1990.

[56] P. Smolensky. On the proper treatment of connectionism. *The Behavioral and Brain Sciences*, 11:1–74, 1988.

[57] M. V. Srinivasan, S. B. Laughlin, and A. Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B*, 216:427–459, 1982.

[58] I. Stewart. *Does God Play Dice?* Penguin, London, 1989.

[59] S. Ullman. The interpretation of structure from motion. *Proc. R. Soc. Lond. B*, 203:405–426, 1979.

[60] T. Winograd and F. Flores. *Understanding Computers and Cognition: A New Foundation for Design.* Addison-Wesley Publishing Co. Inc., 1986.

[61] P. H. Winston. *Artificial Intelligence.* Addison-Wesley Publishing Co. Inc., second edition, 1984.

[62] D. Young. *Nerve Cells and Animal Behaviour.* Cambridge University Press, Cambridge, 1989.

[35] G. Lakoff. *Women, Fire, and dangerous things: What categories reveal about the mind.* University of Chicago Press, 1987.

[36] G. Lakoff. Smolensky, semantics, and the sensorimotor system. *The Behavioral and Brain Sciences*, 11(1):39–40, 1988.

[37] S. B. Laughlin. Form and function in retinal processing. *Trends in Neuroscience*, 10(11):478–483, 1987.

[38] S. R. Lehky and T. J. Sejnowski. Simplifying network models of binocular rivalry and shape-from-shading. In *Methods in Neuronal Modeling: from Synapses to Networks*, pages 361–396. M.I.T. Press — Bradford Books, Cambridge MA, 1989.

[39] R. J. MacGregor. *Neural and Brain Modeling.* Academic Press, 1987.

[40] D. Marr. Artificial intelligence: A personal view. *Artificial Intelligence*, 9:37–48, 1977.

[41] H. R. Maturana, J. Y. Lettvin, W. S. McCulloch, and W. H. Pitts. Anatomy and physiology of vision in the frog. *Journal of General Physiology*, 43:129–175, 1960.

[42] J. L. McClelland and D. E. Rumelhart, editors. *Parallel Distributed Processing. Volume 2: Psychological and biological models.* M.I.T. Press — Bradford Books, Cambridge MA, 1988.

[43] D. K. Pickles. Personal communication, 1989.

[44] D. K. Pickles. Personal communication, 1990.

[45] S. Pinker and A. Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193, 1988.

[46] R. Prakash, E. Solessio, and R. B. Barlow, Jr. Parallel computer model of the *Limulus* lateral eye. *Ann. M. Soc. Neurosci. Abstr.*, 15:1206, 1989.

[47] R. S. Pressman. *Software Engineering: A Practitioners Approach.* McGraw-Hill International, 1982.

[48] F. Ratliff. *Studies on Excitation and Inhibition in the Retina.* Rockerfeller University Press, New York, 1974.

[49] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing, Volume 1: Foundations*, pages 318–362. M.I.T. Press — Bradford Books, Cambridge MA, 1986.

[50] D. E. Rumelhart and J. L. McClelland, editors. *Parallel Distributed Processing, Volume 1: Foundations.* M.I.T. Press — Bradford Books, Cambridge MA, 1986.

[51] T. Scutt. Personal communication, 1990.

[52] T. J. Sejnowski, C. Koch, and P. S. Churchland. Computational neuroscience. *Science*, 241:1299–1306, September 1988.

[53] T. J. Sejnowski and T. Poggio. Series forward. In C. Koch and I. Segev, editors, *Methods in Neuronal Modeling: From Synapses to Networks.* M.I.T. Press — Bradford Books, Cambridge MA, 1989.

[16] D. T. Cliff. Network control for animate vision with nonuniform sampling. CSRP 163, University of Sussex School of Cognitive and Computing Sciences, 1990. Forthcoming.

[17] F. Crick. The recent excitement about neural networks. *Nature*, 337:129–132, January 1989.

[18] D. Dennett. Cognitive wheels: The frame problem of AI. In C. Hookway, editor, *Minds, Machines, and Evolution*. Cambridge University Press, Cambridge, 1984.

[19] H. L. Dreyfus. From micro-worlds to knowledge representation: AI at an impasse. In J. Haugeland, editor, *Mind Design: Philosophy, Psychology, Artificial Intelligence*, pages 161–204. M.I.T. Press — Bradford Books, Cambridge MA, 1981.

[20] J.-P. Ewert. *Neuroethology: An introduction to the neurophysiological fundamentals of behaviour*. Springer-Verlag, Berlin, 1980.

[21] J.-P. Ewert. Neuroethology of releasing mechanisms: Prey-catching in toads. *The Behavioral and Brain Sciences*, 10:337–405, 1987.

[22] S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. CMU–CS–90–100, Carnegie Mellon University School of Computer Science, February 1990.

[23] S. Grossberg. Neuroethology and theoretical neurobiology. *The Behavioral and Brain Sciences*, 7(3):388–390, 1984.

[24] D. M. Guthrie. *Neuroethology: An Introduction*. Blackwell, Oxford, 1980.

[25] D. M. Guthrie, editor. *Aims and Methods in Neuroethology*. Manchester University Press, Manchester, 1987.

[26] S. Harnad. The symbol grounding problem. In *CNLS Conference on Emergent Computation*, Los Alamos, May 1989. Submitted to Physica D.

[27] G. E. Hinton. Connectionist learning procedures. Technical Report CMU-CS 87–115, Carnegie-Mellon University Computer Science Department, Pittsburgh PA, 1987.

[28] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117:500–544, 1952.

[29] I. D. Horswill and R. A. Brooks. Situated vision in a dynamic world: Chasing objects. In *Proceedings AAAI—88*, 1988.

[30] G. Hoyle. Neural mechanisms underlying behaviour of invertebrates. In M. S. Gazzaniga and C. Blakemore, editors, *Handbook of Psychobiology*, pages 3–48. Academic Press, New York, 1975.

[31] G. Hoyle. The scope of neuroethology. *The Behavioral and Brain Sciences*, 7:367–412, 1984.

[32] G. Johnson. *Machinery of the Mind*. Microsoft Press — Tempus Books, Redmond, Washington, 1986.

[33] C. Koch and I. Segev. Introduction. In C. Koch and I. Segev, editors, *Methods in Neuronal Modeling: From Synapses to Networks*, pages 1–8. M.I.T. Press — Bradford Books, Cambridge MA, 1989.

[34] C. Koch and I. Segev, editors. *Methods in Neuronal Modeling: From Synapses to Networks*. M.I.T. Press — Bradford Books, Cambridge MA, 1989.

Acknowledgements

This work was supported by the UK Science and Engineering Research Council, studentship grant #87306795. My ideas for this paper grew over a number of Bacchantic discussions with Paddy Toal and Jaqui Strube. A nascent version of this paper was discussed at the University of Sussex PDP Club, and I thank all participants of that meeting for their interest. Helpful comments on earlier versions of this paper were provided by Harry Barrow, Inman Harvey, David Hogg, David Pickles, Aaron Sloman, and David Young. Inman alerted me to Harnad's work. Many thanks to my supervisors David Hogg and David Young for allowing me the freedom to wander.

References

[1] M. A. Arbib. Levels of modelling of mechanisms of visually guided behaviour. *The Behavioral and Brain Sciences*, 10:407–465, 1987.

[2] M. A. Arbib. Schemas and neural networks for sixth generation computing. *Journal of Parallel and Distributed Computing*, 6:185–216, 1989.

[3] R. B. Barlow, Jr. Personal communication, 1990.

[4] R. B. Barlow, Jr. What the brain tells the eye. *Scientific American*, 262(4):66–71, April 1990.

[5] H. G. Barrow. Personal communication, 1990.

[6] C. C. Bissell. *Control Engineering.* Tutorial guides in electronic engineering: 15. Van Nostrand Reinhold (International) Co. Ltd., London, 1988.

[7] M. A. Boden. Artificial intelligence and animal psychology. CSRP 016, University of Sussex School of Cognitive and Computing Sciences, 1981.

[8] R. A. Brooks. A robust layered control system for a mobile robot. A.I. Memo 864, M.I.T. A.I. Lab, September 1985.

[9] R. A. Brooks. Achieving artificial intelligence through building robots. A.I. Memo 899, M.I.T. A.I. Lab, May 1986.

[10] V. Bruce and P. R. Green. *Visual Perception: physiology, psychology and ecology.* Lawrence Erlbaum Associates, London, 1985.

[11] J. M. Camhi. *Neuroethology: Nerve Cells and the Natural Behaviour of Animals.* Sinauer Associates Inc., Sunderland, Mass., 1984.

[12] F. Clarac. Difficulties and relevance of a neuroethological approach to neurobiology. *The Behavioral and Brain Sciences*, 7(3):383–384, 1984.

[13] A. Clark. Artificial intelligence and the biological factor. CSRP 049, University of Sussex School of Cognitive and Computing Sciences, 1985.

[14] A. Clark. *Microcognition: philosophy, cognitive science, and parallel distributed processing.* M.I.T. Press — Bradford Books, Cambridge MA, 1989.

[15] D. T. Cliff. A closed-environment computational network model of visual processing performed by an airborne insect, 1988. Unpublished D.Phil. Research Proposal. University of Sussex School of Cognitive and Computing Sciences.

Studying animals lowlier than ourselves is not without precedent in other fields. The most notable example is in genetics, where hundreds of thousands of person-years of research have been expended on the fruitfly *Drosophila melanogaster*; with the human genome project commencing only very recently. As John McCarthy (a founder of AI) puts it:

> "It may be that AI is a problem that will fall to brilliance so that all we lack is one Einstein. [but] I think this is one of the difficult sciences like genetics, and it's conceivable that it could take just as long to get to the bottom of it." (McCarthy, quoted by Johnson [32, p.13])

The implication is clear: we should cease to concentrate exclusively on high-level cognitive function and get the basics right first. Such an argument was also proposed by Marr [40], an advocate of low-level studies of human cognition. Marr defended the focus on low-level intellectual function by noting that the lower levels are:

> "...often dismissed with contempt by those who purport to study "higher, more central" problems of intelligence. Our reply to such criticism is that low level problems probably do represent the easier kind, but that is precisely the reason for studying them first. When we have solved a few more, the questions that arise in studying the deeper ones will be clearer to us." [40]

While I agree with Marr's view on which end of the intellectual spectrum it is best to start studying, the difference here is that Marr saw nothing wrong with commencing work at the highest end of the phylogenetic scale. My belief is that following a development path strongly influenced by the evolutionary history of natural intelligence, i.e. starting towards the low end of the phylogenetic scale, will be most profitable.

Even if the study of insects provides us with no data specific to higher animals, it will help clarify the status of the general principles that connectionists search for and which Crick argued against — if the details of the brain *are* so horrendously complicated, then perhaps the answer is not to study simple models of large brains, but to study large models of simple brains.

This shift to an antianthropocentric focus for AI has manifest implications for the connectionist study of language. This is a core cognitive function [56, p.62]. But if the arguments listed above are accepted, then (*pace* the researchers involved) all current connectionist models of language are wildly premature. Language will be best understood as a very high layer in a subsumption architecture: how it interacts with lower layers could be of vital importance, and we should study these lower layers first.

If we are still yet to determine the subsymbolic or neural basis underlying the dance-language of bees, how then are we supposed to study aspects of human language at anything but the most gross level of neuroanatomy (i.e. studies of lesioned patients)? We simply do not know enough.

7 Conclusion

The significance of computational neuroethology is its spanning of all layers of the MacGregor-Lewis stratification. The prospects for computational neuroethology look good; it will never answer all of the questions asked of computational neuroscience, but it will provide "simplifying" network models with well-grounded semantics, and hopefully lead to a consequent rejection of interpretation in favour of observation. Subjectivity will give way to objectivity.

The use of closed-environment simulators, coupled with the rejection of the phylogenetically top-down study of intelligence, hails a new approach. This new approach is evolutionist and antianthropocentric. It focuses attention on the true issues underlying intelligence. Language must wait. Discuss.

... This part of intelligence is where evolution has concentrated its time — it is much harder." [9, p.1]. Clark makes a related point:

> "...intelligence has evolved as a means of satisfying our basic survival requirements. It has *not* been selected for its capacity to achieve the high-level mental feats which so much work in AI is dedicated to modelling. If we *can* perform such feats ...it is only *in virtue* of our being endowed with a set of low-level capacities which just happen to facilitate the higher-level activity.
>
> ...
>
> "The right microworlds to study are not fragments of the sophisticated human achievements, but the less sophisticated achievements of the various animal intelligences, ranged across the phylogenetic tree." [13, pp.4–5; original emphasis]

In particular, Brooks advocates the construction of robotic insects:

> "Insects are not usually thought of as intelligent. However, they are very robust devices. They operate in a dynamic world, carrying out a number of complex tasks... No human-built systems are remotely as reliable... Thus I see insect level behavior as a noble goal for artificial intelligence practitioners. I believe it is closer to the ultimate right track than are the higher level goals now being pursued." [9, p.7].

Similarly evolutionist arguments have been proposed in neuroethology. Probably the most passionate advocate of restricting study to arthropods (the animal class to which insects belong) was Graham Hoyle:

> "Many neurophysiologists... express no, or very little interest in invertebrates. Their goal is to understand the higher mammalian nervous system, period. ... The invertebrates have nervous systems which will certainly be understood first in cellular and connectivity terms... by definition, the general principles must be those features which vertebrates possess in common with invertebrates... The study of specifically human nervous systems is an applied, not a general, science." [30, p.17].

From the point of view of neural modelling, a fundamental advantage lies in insect neuroarchitecture. There is ample data in the biology literature demonstrating that insect (and other arthropod) nervous systems are constructed with a particularly economical use of neurons. For example, insect muscles are generally innervated with no more than six neurons, while mammals may use hundreds of nerve cells to perform the same excitatory function [62, pp.3–4]. Similarly there is persuasive evidence [37, 57] that insect nervous systems are very finely tuned to maximise information capacity in the presence of noise. This contrasts with the commonly held view that mammalian nervous systems transmit identical information over many parallel independent channels, which are subsequently integrated to reduce the disruptive effects of noise.

As Young notes [62, p.4], the smaller number of nerve cells involved does not necessarily indicate that the neural principles of operation are simpler in insects than in higher animals. However, it does mean that a computational network simulation of insect nervous function is closer in architectural terms to biological reality than any corresponding model of mammalian or primate function, by several orders of magnitude.

So, perhaps the best approach to studying intelligence is, phylogenetically speaking, bottom-up rather than top-down. All previous work in cognitive modelling has focused on advanced animals, near the top of the phylogenetic hierarchy, i.e. humans (and, in vision research, other mammals). Yet the only case for such anthropocentricity is an *a priori* one. Studying intelligence by modelling insects might seem counterintuitive, but then again it might yield some useful results. Perhaps we have been missing something.

Thus the central ideas in the Brooksian philosophy are: that the control of a mobile robot can be viewed in terms of behaviours rather than functional modules; that the subsumption architecture allows for incremental construction and debugging of complex mobile robot control systems; that a low-bandwidth network of loosely coupled simple asynchronous processors can perform the required computation; and that there is no need for a central control module: "The control system can be viewed as a system of agents each busy with their own solipsist world." [8, p.19].

Brooks claims [9, pp.6–7] two main advantages for systems built according to his principles: parallel control paths and parallel behaviour generation. The presence of many parallel control paths through the system means that the most relevant behaviour is activated for a given situation: there is no longer a "weakest link" on which the system's entire performance depends. When more than one behaviour is appropriate in a given situation, parallel behaviour generation occurs: there is no central bottleneck of control, and desirable redundancy and robustness result from having a choice of behaviours available.

The intention then in closed-environment simulator computational neuroethology is to create a subsumption architecture where the modules forming each layer are not formal (symbolic) finite state automata, but small artificial neural networks. A network model constructed according to a subsumption architecture is an explicit recognition of the inhomogenous nature of natural nervous systems.

Two guiding principles in Brooks's work that are particularly relevant to such an approach are:

> "Complex (and useful) behavior need not necessarily be a product of an extremely complex control system. Rather, complex behaviour may simply be the reflection of a complex environment [[54]]. It may be an observer who ascribes complexity to an organism — not necessarily its designer." [8, p.3]

> "For robustness sake the robot must be able to perform when one or more of its sensors fails or starts giving erroneous readings. Recovery should be quick. This implies that built-in self calibration must be occurring at all times. If it is good enough to achieve our goals then it will necessarily be good enough to eliminate the need for external calibration steps. To force the issue we do not incorporate any explicit calibration steps for our robot. Rather we try to make all processing steps self calibrating." [8, p.4]

More recently, Horswill and Brooks [29] have described a robotic cart driven by a layered control system. This cart has a top-level objective of tracking (pursuing) objects presented to it. The objective is successfully achieved in real-time using a comparatively low-resolution vision system as it's sole means of input. The recently devised cascade-correlation neural network learning technique [22] bears a conceptual similarity to subsumption architecture.

6.3.2 AI: Artificial Intelligence or Artificial Insects?

Moreover, Brooks argues that there are severe limitations on the generally accepted methodologies of artificial intelligence, and that the currently accepted conceptual decompositions and static representations are wrong. He argues for a shift to process-based modelling, contending that: "...mobility, acute vision, and the ability to carry out survival related tasks in a dynamic environment provide a necessary basis for the development of true intelligence." [9, p.2]. Brooks's argument is based in part on the evolution of intelligent beings: that intellectual capabilities such as writing and 'expert' knowledge are, on the evolutionary timescale, very recent developments indicates to him that the creation of artificial entities truly possessing such abilities will be relatively straightforward "...once the essence of being and reacting are available.

cohesive,[10] modules. Such systems are flawed in that they are only as strong as their weakest link, there is normally a central locus of control (either declarative or procedural: either a data bottleneck or a means of propagating errors) [9, p.6], and all modules need to be in place before the system is truly operative.

Brooks proposes that the primary decomposition of the control problem should be into *task achieving behaviours*. That is, he advocates dividing the desired intelligent behaviour of a system into a collection of simpler behaviours; behaviourally complex systems are thus composed from a number of computational systems achieving simpler behaviours. The intention is that the simpler systems should be independent, but some degree of overlap is likely in practice.

Brooks organises the behavioural decomposition on the basis of two concepts: task achievement and competence levels. Task achievement dictates that: "Each behavior should achieve some task. I.e., there should be some observable phenomenon in the total behavior of the system which can be used by an outside observer to say whether the particular sub-behavior is operating successfully." [9, p.6]. Competence levels specify overall system performance in a rather informal manner: "A set of task achieving behaviors together provide the robot with some level of competence. They should be designed so that as new task achieving behaviors are added to the system, the level of competence increases." [*ibid.*]. Brooks further discusses a layered control system:

> "Layers of control system are built to let the robot operate at increasing levels of competence. Layers are made up of asynchronous modules which communicate over low bandwidth channels. Each module is an instance of a fairly simple computational machine. Higher level layers can subsume the roles of lower levels by suppressing their outputs. However, lower levels continue to function as higher levels are added. The result is a robust and flexible robot control system." [8, abstract]

For example, the first four levels of competence of the robot constructed by Brooks are [8, pp.6–7]:

0. Avoid contact with objects (whether the objects move or are stationary).
1. Wander aimlessly around without hitting things.
2. "Explore" the world by seeing places in the distance which look reachable and heading for them.
3. Build a map of the environment and plan routes from one place to another.

These layers are implemented to form what Brooks [8, p.7] calls a *subsumption architecture*: initially, the complete robot control system is constructed to achieve level 0 competence — this is referred to as the zeroth layer control system; once it is completely debugged it is never subsequently altered. Next, the first layer control system is constructed: it can receive input from the layer 0 system, and additionally it can suppress, or *subsume*, the output of the zeroth layer. The zeroth and first layers together implement level 1 competence. Similarly, higher layers are constructed to realise higher levels of competence.

A particular layer always runs unaware of the activities of all higher layers, any of which might interfere with its datapaths. Each level of competence includes as a subset each earlier level of competence [8, p.7], so successive layers can be viewed as representing increasingly more sophisticated and constrained classes of allowable behaviour. In Brooks's original robot, each layer is built from a number of asynchronous modules, and each module is a finite state machine [8, p.9].

[10] Cohesion is sometimes referred to as *binding* in the literature (e.g. [47, pp.158–164]).

Brooksian sense, then the model ceases to demand pure *interpretation*; most results are *observed*. That is, discussion of the model network ceases to be solely *subjective* and becomes much more *objective*. Soft data is replaced by hard data.

This is not to say that *all* networks without sensorimotor linkage are necessarily doomed to purely subjective interpretation. Nevertheless, too many simplifying connectionist models are presented as evidence for (as yet) unfalsifiable claims about the biology they are purported to represent. Analysis and interpretation will still be required in closed-environment computational neuroethology, but it is possible to talk of computational neuroethology as a falsifiable endeavour. This is not before time. There is a significant amount of connectionist literature which treats the methodology as some form of occult lore: little understood and more powerful than those who practice it.[9] The sooner this 'neuromancy' is replaced by something approaching 'neurosmithy', the better.

6.3 Finale: Implications

The implications of computational neuroethology for cognitive science are, I believe, more fundamental than simply tightening up a slack discipline. I briefly outline some thoughts below, but the credit is due to others: the inspiration comes mainly from past work in robotics by Rodney Brooks and in neureothology by Graham Hoyle.

The need for a behavioural linkage, and its satisfaction using a closed-environment simulation system, raises a supplementary issue: if there is so much emphasis placed on the preservation of the data-path from sensory input to motor output, how is it possible to model high-level cognitive functions where there is believed to be significant processing performed between input and output? Surely the sequence of functional units between input and output in complex tasks such as speech understanding or three-dimensional reasoning from vision is so long as to make the modelling task prohibitively complex?

If this problem cannot be resolved, at least in principle, then closed-environment simulation is only useful for a very small class of problems: problems which traditionally have been peripheral to cognitive modelling and artificial intelligence.

Fortunately, help is at hand. This problem is resolved by careful attention to two assumptions implicit in its formulation: first, that a sequential functional decomposition of cognitive processes is the best approach; second, that what have traditionally been 'core' topics in cognitive modelling really are core topics, i.e. that problems amenable to closed-environment simulation *really* are peripheral issue. Both of these assumptions have been questioned and found wanting by Brooks.

6.3.1 Subsumption Architecture

Brooks [8] has developed an approach to the design of control architectures for mobile robots operating in the real real world, which can readily be adopted for use in the simulated real worlds of closed-environment simulators. The key to Brooks's approach is the concept of a *layered* control system, which leads to control architectures and strategies radically different from convention.

The conventional approach to constructing a mobile robot control system is to decompose the problem into a set of *functional units*, creating an essentially sequential linear datapath which starts at the sensory transducers and ends at the motor actuators. The software structure of a control system based on such functional decomposition typically has highly coupled, logically

[9] For example, at a major European conference in 1989, only one connectionist paper was accepted. This paper employed back-propagation to create a network. Both the author and the referees missed the fact that the network was underconstrained and was merely performing table-lookup. Simple algebra from network statistics explicit in the paper would have revealed this. No names, no pack-drill.

a closed-environment model. No humans are required.[7] The feedback properties of the environment have long been recognised by (neuro)ethologists: e.g. it is referred to as "reafference" in the efference copy hypothesis of insect optomotor movements, commonly attributed to von Holst and Mittelstaedt in the mid 1950's, but first suggested by von Helmoltz in 1925 (see [11, p.342]).

Closed-environment simulator systems already exist, albeit not in neuroethology but in e.g. the domain of military and commercial flight simulators. Here the neural network is real (and human) but that is not material: motor commands are issued (to the cockpit control devices such as joystick and engine throttle) and the simulator performs the computations necessary on a model environment to generate the sensory input (e.g. dynamic 3-D graphics and sound, movement of the cockpit, etc.) which provides sensorimotor feedback.

Closed-environment simulators for neuroethology can be fashioned in much the same manner. The simulation of the medium thus involves *two* models: the model network and the model environment.[8] The complexity of the simulation is dependent on the theory implemented by the simulator. This is an important issue, pointed out to me by Pickles [43]: in order to simulate something, you should first have a theory of that thing, because the simulation is an embodiment of the theory, even if the only statement of the theory is implicitly in the specification of the simulator.

This has bearing because the use of simulated environments is open to the accusation of falling into the micro-world trap, which dogged AI through the early 1970's (see [19] for a discussion and pointers to the micro-world literature). Closed-environment simulators are not subject to micro-world problems because they are not "toy" systems reliant on human interpretation. They are real-world models for a certain definition of real. Consider:

> "There is great methodological danger in tackling AI problems in toy worlds. The definition of "toy" here includes all worlds where it is not up to the AI system itself to do all the understanding of the world itself without relying on a human interpreter. Likewise a world is a "toy" world if the AI system is not responsible for carrying out its actions in the world without a human agent to interpret its responses. Such requirements on an AI system thus force it to be part of a robot system acting in a *real world* for some definition of real. This is a much stronger definition of real world than is normally used." Brooks [9, p.5, original emphasis].

Even under such a strong definition of real world, it is possible to work on real-world problems using closed-environment simulators rather than robotics (thus ignoring some of the more problematic engineering issues of sensorimotor transduction). An existence proof is the SYCO simulator of visual processing in a hoverfly [16].

The advantage in the use of simulated "real" worlds rather than *the* real world (i.e. a robotics approach) is that the real real world has many superfluous degrees of freedom: many experiments in e.g. visual neuroethology take great care to minimise the dimensionality of the experimental animal's world by creating visually impoverished environments (such as striped optomotor cylinders: e.g. [10, p.211]). Such precautions are also necessary, at least during development and testing, in perceptual robotics; but in simulated "real" worlds *identical* conditions can be recreated as many times as is required.

The closed-environment simulator approach thus provides a behavioural linkage for a simulated neural network. It grounds the semantics of the network in the semantics of the simulated environment. And this is the fundamental point: if the simulated environment is "real" in the

[7] Of course, humans are needed to set up the model and its interfaces, but we're not that interested in where the model comes from: it's what it does that counts.

[8] Strictly, according to Maturana, it is one model (the medium) with two components: the network and everything-but-the-network.

It is in this respect that connectionist computational neuroscience has most to learn from neuroethology, and it is this belief in the importance of behavioural linkage that most distinguishes computational neuroethology. So, computational neuroethology replaces the *in vacuo* approach of connectionism with a (simulated) *in vivo* approach; and in doing so, the semantics of the model are automatically grounded.

Thus, computational neuroethology can be provisionally defined as being the study of neuroethology using the techniques of computational neuroscience. This definition intentionally admits many classes of model, but the significant aspect is the increased attention to the environment that the neural entity is a component of. The vertical nature of neuroethology precludes restriction purely to "simplifying" (connectionist) models: modelling techniques from realistic computational neuroscience should also be applicable to computational neuroethology.

Furthermore, this broad definition is noncommittal on the interpretation of the word "computational": a strict interpretation would allow only neuroethological models that focus on behaviour as a result of computation; but the vertical nature of neuroethology surely indicates that models with no *direct* reference to computation should still be of interest, and the Maturanian anti-pipeline argument outlined above additionally indicates that sole focus on information processing may omit important principles. So "computational" can simply imply a reliance on computerised experimental techniques, as in e.g. computational physics.

And that, for the time being, is the working definition of computational neuroethology. It is a broad definition, but this is a provisional manifesto, and ensuing debate may refine the definition. What is required is guidance on *how* the connectionist model should be connected to a sensorimotor system, thereby granting the required behavioural linkage. This is a subissue of the field, and is discussed at length in the next section.

As I mentioned above, the term "computational neuroethology" was coined to describe my work in studying visual control of insect flight [16]. There is plenty of other research that can be classed as work in computational neuroethology: a few examples follow. Robert Barlow has agreed [3] that the work of his research group on a Connection Machine model of the compound eye of *Limulus* [4, 46] is a kind of computational neuroethology. Tom Scutt uses the term to describe his modelling of neurons in the *Aplysia* network [51]. Also I would classify Michael Arbib's work on the computational frog *Rana computatrix* as computational neuroethology, although he himself has described his work as "(theoretical) neuroethology" [31, p.381] and "connectionism; neural modelling, neuroethology" [1, keywords, p.407].

6.2 The behavioural link

This section discusses my personal attitude towards the issue of providing connectionist models with a behavioural linkage. As such, it is only one of many possible approaches within computational neuroethology.

The need for a behavioural linkage is satisfied by incorporation in the model of the importance of the environment within which the neural entity is a component. The methods of connectionist computational neuroscience can be adapted to computational neuroethology in a fairly straightforward way: by embedding the network model within a *closed-environment* simulator. A closed environment simulator is one which provides a datapath that models the external feedback loop provided by the environment.

That is, closed-environment models eliminate any "humans in the loop": current connectionist models have a data-path that resembles the architecture of an open-loop control system (e.g. [6, p.4]); humans are responsible for feeding data (and meaning) in at the front end and collecting output data (and assigning meaning) at the back end. Closed-environment models are closer in nature to closed-loop control systems (e.g. [6, p.5]): the feedback loop in a closed-loop controller corresponds to the sensorimotor feedback provided by the simulated environment in

ad hominem, but Smolensky's views are, I believe, representative of the field. Acceptance of *in vacuo* modelling is pandemic within connectionism.

5.5 Summary

Several criticisms of connectionism have been discussed. It has been argued that the connectionist paradigm is biologically *in vacuo* and in this sense is no advance on the symbolic paradigm: connectionism has acted merely as a palliative for several of the maladies of symbolism. The solution to this problem is linking the model neural network to the external world via a sensorimotor system, thus grounding the symbols in the model. Such an approach is more in line with the philosophy of the Maturana school. Such an approach is embodied in computational neuroethology.

6 Computational Neuroethology

6.1 Towards a definition

Computational Neuroethology (as far as I am aware, I originated this term, in [15]) is proposed here as a method of eliminating the *ad hoc* semantics of contemporary connectionist models. The definition of the term "computational neuroethology" builds on the definition of "neuroethology", given below.

Neuroethology (e.g. [24, 20, 11, 25, 62]) is a young discipline within the biological sciences: it is where ethology (the study of behaviour) meets neuroscience. Its youth puts it in a situation not dissimilar to cognitive science; its precise aims and methods are still a subject of debate (see e.g. [31, 25]). Put most simply, neuroethology is the study of the neural mechanisms of animal behaviour [31, p.384]. More specifically, the goal of neuroethology is to relate behaviour to activity within interconnected groups of nerve cells, where behaviour is defined as patterns of movement coordinated in space and time (this definition is due to Ewert [31, p.378, p.381]). As Clarac remarks, "By definition, neuroethology is a "vertical" science, whose main interest is to link the results obtained from many different levels of complexity in the nervous system." [12, p.384]. That is, neuroethology spans all layers of the MacGregor-Lewis hierarchy, and thus (according to MacGregor [39, p.8]) neuroethological models are the most broadly significant models in neuroscience.

So, if neuroscience can be said to take a "local" view of an organism, neuroethology takes a "global" view, verging on the holistic. Many researchers believe that it is important to follow a *comparative* neuroethological approach, i.e. exploring the neural principles in animals of different ecological and behavioural adaptations. It is believed that such an interspecies approach may contribute to the formalising of general rules and concepts, providing insight into general neurophysiological principles underlying behaviour [31, p.387].

The relevance of neuroethology to the *in vacuo* problem in connectionism is neatly captured by Grossberg:

> "Neuroethology teaches us that neural circuits are organised to generate adaptive goal-oriented[6] behaviors. Without a behavioral linkage, no amount of superb neurophysiological experimentation can lead to an understanding of brain design, because this type of work, in isolation, does not probe the functional level on which an organism's behavioral success is defined." [23, p.389].

[6] "goal" is used here in the teleological sense rather than the intentional sense.

The essence of Lakoff's argument is thus that if neurons are appropriately located relative to the sensorimotor system then activation patterns over a network of neurons are meaningful *in themselves*. That is, the activation patterns do not have to be "given meaning" in the same way that symbol-strings do. But surely any connectionist model that is *not* connected to a sensorimotor system *has* to be "given meaning" in much the same way as symbolic systems do: whether you are giving meaning to one discrete symbol, or to a configuration of activations over a vector of input or output units, you still have to associate these (sub)symbols with things in the world; you are still the source of meaning.

Lakoff continues:

> "In a full-blown connectionist theory of mind activation patterns over neurons are meaningful in themselves by virtue of what they are connected to. The intractable problem of assigning meanings to symbols does not arise here.
>
> "It is also important to remember that the isolated models connectionists build to study the properties of networks are not full-blown connectionist theories of mind. They vastly over-simplify, or totally ignore, sensorimotor input and output, assuming that, for the purpose of the study at hand, one can just as well use feature names, to which the model-builders must assign meanings. This is a crucial difference between isolated models and a full-blown theory.
>
> ...
>
> "...it is vital to bear in mind that a full-blown connectionist theory of mind is a lot more than just an information-processing system." [36, p.40]

The background to Lakoff's argument is presented in [35]. The need for neural grounding is also argued for, albeit less forcefully, by Mortenson [56, pp.44–45]. Harnad [26] develops a very similar argument to Lakoff's, and proposes a method for grounding *symbolic* models in sensorimotor systems, via connectionist processes.

The argument that the linking of neural models to a sensorimotor system provides an automatic grounding for the semantics of the model is important, so some further clarification is justified. The precise notion of 'meaning' is a deep philosophical problem, but for the present purposes[5] we can take Lakoff's 'meaning' to be *the content of the internal representations* posited by the model [44]. Then I'm interpreting Lakoff's argument as:

> The representational contents of the constituent states of a computational process not connected to a sensorimotor system are dependent on the intentions of the person who designs and observes the process. If the programmer says that 00100011 represents *Dog* then it represents *Dog* and that's all there is to it. On the other hand the contents of the states of a computational process which *is* connected to a sensorimotor system are, in a sense, *objective*. They do not depend on the intentions of a designer or observer. They depend, rather, on the role that they have in the sensorimotor system. [44]

That is, saying that the semantics of the system are automatically grounded is another way of saying that the system's representational content is dependent on being embedded in a sensorimotor system.

So, pursuing this approach prevents theorists from inventing the meaning of patterns of activity (as they are free to while the models remain in their current isolated status). Smolensky's reply to this ("[it] seems an important philosophical point, but one that cannot really do any modelling work until the gap is bridged (at least partly) between the subconceptual [i.e. connectionist] and neural levels..." [56, p.66]) is distinctly *mezza voce*. Again, the attack here isn't

[5] The treatment of semantics in this paper is rather simplistic, but necessarily so in order to avert an undesirable digression into the meaning of "meaning". See [55] for a more complete treatment of machine semantics.

> "First, isolate an interesting human achievement, story understanding, say. Second, find the best way you can of getting a conventional Von Neumann [*sic*] computer to simulate some allegedly central chunk of the input-output profile associated with human performance of the task. Finally, hope that the program so devised will be at least a clue to the form of a good psychological theory of how human beings in fact manage to perform the task in question." [14, p.61].

Clark argues that this approach is "...as biologically implausible as it is philosophically unsatisfactory. No serious study of mind (...) can, I believe, be conducted in the kind of biological vacuum to which cognitive scientists have become accustomed." [*ibid.*].

I do not take issue with the accusation that the classical-cognitivist approach is biologically *in vacuo*. But I have grave doubts that current connectionist computational neuroscience offers any real improvement. In fact I see no difference in principle between the current connectionist approach and the classical-cognitivist: connectionism just involves a switch to biological terminology. As illustration, compare the Clark quote given above with this slightly altered version:

> First, isolate an interesting nervous system achievement, visual orientation detection, say. Second, find the best way you can of getting a sexy connectionist computer to simulate some allegedly central chunk of the input-output profile associated with neural performance of the task. Finally, hope that the program so devised will be at least a clue to the form of a good biological theory of how neural networks in fact manage to perform the task in question.

This captures the spirit of most connectionist research with which I am familiar. The tune may have changed, but the song remains the same.

5.4 Grounding the semantics

This final criticism of connectionism continues the theme of the *in vacuo* problem, demonstrating the full extent to which connectionism exists within a biological vacuum. This criticism is Lakoff's critique of the semantics of connectionist systems; a similar argument has been proposed by Harnad [26]. Lakoff [36] criticises Smolensky's [56] account of the connectionist paradigm as making a huge omission in ignoring the body. He states:

> "The neural networks in the body do not exist in isolation; they are connected to the sensorimotor system. For example, the neurons in a topographic map of the retina are not just firing in isolation for the hell of it. They are firing in response to retinal input, which in turn is dependent on what is in front of one's eyes. An activation pattern in the topographic map of the retina is therefore not merely a meaningless mathematical object in some dynamical system; it is *meaningful*. A different activation pattern over those neurons would mean something different. One cannot just arbitrarily assign meaning to activation patterns over neural networks that are connected to the sensorimotor system. The nature of the hookup to the body will make such an activation pattern meaningful and play a role in fixing its meaning.
>
> "Compare this, for example, with a string of symbols in a ...computer program. The symbols are not meaningful in themselves. They have to be "given meaning" by being associated with things in the world. If the symbols are to stand for categories, those symbols must be given meanings by being associated with categories that are out there in the world." [36, p.39]

5.3.1 Micro-worlds, again

The first example concerns the micro-world trap in symbolic cognitive modelling. This is about previous work in the symbolic paradigm which proposed to have created systems which 'understood' concepts or which were claimed to exhibit learning or 'discovery' of new concepts. There is a lot of such work, and a comprehensive literature review is out of the question here: this is an eclectic overview of the argument.

Essentially, the criticism of such models was that all the 'understanding' was actually being performed by the programmers, because in creating the system it was necessary to create a micro-world for the system to work on [19]. Such toy worlds implicitly incorporated a vast amount of pre-processing. For example, one well-known system was claimed to draw analogies from Shakespeare's plays: actually it was 'spoon-fed' a minimal prećis of *Macbeth*, which is about 100 words long when expressed in English [61, p.413] and requires twelve assertions in a declarative representation [60, p.122]. The problem is that such programs have a claim to understanding "...which is based on the fact that the linguistic and experiential domains the programmer is trying to represent are complex and call for a broad range of human understanding. ...the program actually operates within a narrowed micro-world that reflects the blindness of that representation." [60, p.123].

Similarly, systems which were purported to learn or discover new concepts make their discoveries by "...working on data presented in notational formats that represent the fruits of centuries of human labor. Manipulating these representations could be the tip of the iceberg; creating them and understanding them may constitute the unseen bulk." [14, p.14].

One of the appeals of the connectionist approach listed by its proponents is the availability of autonomous learning procedures for tuning the weights in the network, thereby 'programming themselves' to perform the task at hand. At first glance, the supposed autonomy (no human intervention is required to actually alter the weights) and the distributed nature of the resultant representation both would seem to indicate that connectionism does not suffer from microworld problems.

Such learning techniques are conventionally classified into three broad classes: supervised, reinforcement, or unsupervised; examples of all classes are given in [27]. The three classes all share the need for a source of input vectors — specifications of the activities of the input units of the network: they differ in the degree to which they require additional information in order to correctly tune the weights. And there's the rub. However a network is expected to learn, it requires a source of input vectors which, unless the network is involved in the first stages of vision or hearing, has to be prepared "off-line", invariably by humans rather than by other connectionist networks. That is, connectionist models which rely on pre-processed information are susceptible to the same problems as beset the micro-world studies of the symbolic paradigm. Smolensky acknowledges this problem. He notes [56, pp.7–8] how important the choice of representation for input (micro)features is, and notes that a model's performance depends crucially on the input and output representations chosen by the modeller.

Probably the most celebrated example of this is the Rumelhart and McClelland verb past-tense learning system [42, pp.216–271]: its 'wickelfeature' representation scheme (among other things) has been attacked, most notably in the critique by Prince and Pinker [45, 56, pp.46–47] where the model is described as a fairy-tale account of the actual cognitive process.

5.3.2 Sustaining the vacuum

The second example concerns the belief of many practitioners and observers of connectionism that it is somehow methodologically 'closer' to biology than the preceding symbolic paradigm was. For instance, Clark refers to the old symbolic approach as the classical-cognitivist approach, and characterises it thus:

This maths-driven nature of connectionist architecture is acknowledged by Smolensky: "In the drive for more computational power, architectural decisions seem to be driven more and more by mathematical considerations and less and less by neural ones." [56, p.9].

So, the view (implicit in connectionism) of nervous systems as input/output pipeline devices is mistaken and should be avoided. Such an opinion was also expressed by Hoyle:

> "I am ready to take seriously any model which treats nervous systems as devices for the intrinsic generation of behaviour as their primary purpose. Inputs which promote general or specific behaviours, or which modify ongoing behaviour, or which are anticipated and incorporated into the generation of specific behaviours, are superimposed upon an intrinsic capability. We must reject all models (they are the majority) that treat nervous systems as if they were input/output devices." [31, p.408]

Furthermore, Maturana's philosophy admits no difference, from the standpoint of the nervous system, between a 'real' stimulus and a hallucinatory one. For instance, when light hits the retina, it triggers chemical changes in the photoreceptors and subsequent interneurons, i.e. structural changes occur in the nervous system. However, if a chemical irritant is injected onto the retina causing the same structural changes, then there is no way the nervous system can distinguish between the two forms of perturbation. Perception is an oversimplifying concept in this context because to talk of 'perception' of the light is to admit talk of 'perception' of the irritant: the question of whether the percept of light was a hallucination or was really perceived makes no sense at the neurophysiological level [60, pp.42–43].

As Winograd and Flores note: "Maturana argues that all activity of the nervous system is best understood in this way. The focus should be on the interactions within the system as a whole, not on the structure of perturbations. The perturbations do not determine what happens in the nervous system, but merely trigger changes of state. It is the structure of the perturbed system that ...*specifies* what structural configurations of the medium can perturb it." [60, pp.42–43, original emphasis]. The term 'medium' has a special meaning here: it refers to the space in which an organism exists; an entity exists not as a separate object inside a medium, but as a part of it. That is, there is no modular separation between an entity and its 'environment' [60, p.43]. That an entity is defined only in terms of its surrounding environment is an issue ignored by virtually all current connectionist models. Most treat the portion of the nervous system under study as capable of being modelled *in vacuo*. At best, connectionist models are extremely poor approximations to *in vitro* studies.

5.3 Microscope Envy

Unfortunately, the connectionist disregard for past work in biology does not stop at the philosophical level of Maturana. Lehnert's diagnosis of connectionists as closet physicists suffering from theorem envy was noted above. I suggest that the complementary case also applies: some connectionists are physicists or mathematicians who are closet biologists suffering from microscope envy. Specifically, it appears that one concrete achievement of the connectionist paradigm has been to recast old problems of the symbolic paradigm at a new, biological, level.

This is a problem rooted in the cavalier attitude most proponents of connectionism adopt towards biological reality. For example, The phrase "neuron-like" (as in "a network of neuron-like units...") peppers the literature, to the point where "neuron-like" apparently means "mapping many inputs onto one output through a nonlinearity". In which case the euclidian n-space distance operator $D_n(\mathbf{p},\mathbf{q}) \equiv \sqrt{(\sum_{i=1}^{n}(p_i - q_i)^2)}$ is neuron-like. I'll limit my discussion of old problems recast to two examples.

is extreme, but no more extreme than the *a priori* assumption made by mathematically oriented researchers that Type 1 theories *are* available for whatever aspect of brain function they turn their attention to. As Boden noted, in attempting to understand Nature we must take care not to be seduced by mathematics at the cost of considerations of reality. Maths is useful, but not that useful.

5.2 Biological studies of cognition

The preoccupation with elegant mathematics has led, I believe, to connectionists ignoring a significant school of thought in neuroscience: that of Humberto Maturana and his colleagues. Maturana's work is concerned with understanding how phenomena such as cognition and language are rooted in biological processes. He was involved in pioneering studies of the neurophysiology of vision (e.g. [41]), later developing theories of the organisation of living systems and of language and cognition. Maturana's work is challenging, using a large specialised vocabulary, and a full review of his work is way beyond the scope of this paper: it will suffice here merely to nod in his direction. The notes that follow are based on an overview of his work by Winograd and Flores [60, pp.38–53].

The aspect of Maturana's work which I feel connectionism has most to gain from concerns the closure of the nervous system [60, pp.41–44]. Essentially Maturana's argument is that perception should be studied by examining the properties of the nervous system not as a filter on the mapping of reality, but as a generator of phenomena.

This is illustrated by an example from colour perception. If an object is illuminated from opposing sides by two lights, one white and the other red, then it casts two shadows. On the background of pink light, one of the shadows appears red and the other appears *green*. But this is in the absence of light with significant energy in the wavelengths around 530–560nm which is usually called green. The argument of Maturana and his colleagues is that perception of the green shadow produces the same neural activity patterns as would green light. Thus: "the presence of 'green' for the nervous system is not a simple correlate of the presence of certain wavelengths of light, but the result of a complex pattern of relative activity among different neurons." [60, p.41].

Maturana realised that the study of colour vision should not be envisaged as studying the mapping of a colourful world onto the nervous system, but rather as understanding the participation of the retina or nervous system in the generation of the colour space of the observer. As Winograd and Flores relate:

> "Maturana describes the nervous system as a closed network of interacting neurons such that any change in the state of relative activity of a collection of neurons leads to a change in the state of relative activity of other or the same collection of neurons. From this standpoint, the nervous system does not have 'inputs' and 'outputs'. It can be *perturbed* by structural changes in the network itself, and this will affect its activity, but the sequence of states of the system is generated by relations of neuronal activity, as determined by its structure." [60, p.42, original emphasis]

The lesson to be learned from this is that the nervous system should not be treated as an input-output device. But most connectionist models do exactly that. That is, they treat the neural function to be modelled as being implemented on a 'pipeline' processor. The most likely reason for this is the mathematical tractability of "feed-forward" network models, where each unit sends output only to units in subsequent layers in the network (i.e. later stages of processing). For instance, the back-propagation network learning algorithm [49], the subject of much attention in the literature, operates only on feed-forward networks and is defined solely in terms of the input-output profile of the function at hand.

rendously complicated that no good will come of cramming one's head with that sort of information." [17, p.132]

Crick's ontological objection is rooted in his methodological criticism. He argues that the search for mathematical expression is highly questionable because there is no guarantee of deep general principles being embodied in the functioning of the brain. He offers the genetic code as a good example of a complex natural system which is not easily characterised by a small set of general principles: the brain may achieve its aims using a "series of slick tricks" [17, p.132]; and he contends that further research should be aimed at resolving this issue:

> "Learning about neurons, their behaviour and their connections, will not by itself solve our problems, but will at least suggest the sort of answer to look for and can be used, often rather decisively, to disprove false theories." [17, p.132]

Crick's doubts over the existence of deep principles of brain function can be phrased in terms of Marr's division of studies of biological information processing into Type 1 and Type 2 theories [40]. Briefly, Marr defines a Type 1 theory as one which accounts for a problem by decomposing it into a set of computations which must be performed: Type 1 theories are little concerned with algorithmic details of implementing the computation; they are prealgorithmic in that they focus on *what* any algorithms should be doing, and such information is captured in a computational theory of the function under study. As Marr notes, "The fly in the ointment is that while many problems of biological information processing have a Type 1 theory, there is no reason why they should all have. This can happen when a problem is solved by the simultaneous action of a considerable number of processes, *whose interaction is its own simplest description,* and I shall refer to such a situation as a *Type 2* theory." [40, original emphasis]. Marr's example of a candidate for a Type 2 theory is the problem of predicting how a protein will fold. Marr points out that there is a spectrum of possibilities between Type 1 and Type 2 theories: the distinction is not a pure dichotomy.

So, in Marr's terms, Crick's objection is that current neural network models are founded on the premise that Type 1 theories *are* available for the aspects of brain function under study, and he believes that it is likely that only Type 2 theories will be useful. In this phrasing, Crick's views appear somewhat extreme, and some researchers [5] view his critique as a caricature of the field.

In the interest of balance, it is worthwhile to note that such focus on mathematics at the expense of other (biological) considerations is a feature also found in the symbolic paradigm. For instance, Boden [7] criticised Ullman's [59] work in low-level vision, where an attempt to identify the minimal computational processes required to explain motion perception led to ethologically implausible hypotheses:

> "What is ethologically implausible about Ullman's hypotheses is ...that they assume the perception of *rigid* objects to be basic, while perception of non-rigid movement is taken to be a more complex special case. *Mathematically*, of course, the perception of non-rigid motion is more complex; but this does not prove that it is *biologically* secondary to the perception of rigid objects. ...Admittedly, a robot could be provided with an Ullmanesque capacity to perceive rigid objects in motion; but whether any creature on the phylogenetic scale employs such visual mechanisms is another question." [7, p.10, original emphasis]

Nevertheless, Crick's objections cannot be ignored. The focus on elegant mathematics can be read as an axiomatic acceptance of the existence of Type 1 theories "waiting to be discovered"; and he rightly draws attention to the contrasting view that brain function will only ever usefully be described by Type 2 theories. In my opinion, his apparent dismissal of the Type 1 approach

and *implementationalism*. Briefly, the implementationalist view is that symbolic processing models of cognitive processes are exactly correct accounts, and that connectionist (subsymbolic) accounts are merely implementations of the symbolic processes. That is, they are still exactly correct but they are irrelevant because they contribute nothing new. At the other extreme, the eliminativist view is that symbolic models are nothing but folklore. Smolensky partitions the eliminativist view into two factions: the most extreme is the neural-eliminativist, which additionally advocates that the subsymbolic level of connectionist modelling is nonexistent, while the connectionist-eliminativist maintains that subsymbolic models are in fact exactly correct.

Smolensky argues that the proper treatment of connectionism is to place it slightly to the implementationalist side of connectionist-eliminativism; a position he describes as *limitivist*. The limitivist view still holds that subsymbolic models are exactly correct as cognitive models, but it concedes that symbolic models are approximately correct.

5 Out of the frying pan, into the fire?

Connectionism has not been received with universal acclaim. Criticisms have been made both of specific models and of the methodology as a whole. Presented below are summaries of critiques by some other authors, along with some criticisms of my own. Again, this paper is only interested in (attacks on) connectionism *qua* simplifying computational neuroscience. I am aware that some criticisms have been replied to by proponents of connectionism, but a full review of the debate would be an unwelcome digression here. The four criticisms made below are those which I feel connectionism has most consistently failed to acknowledge or answer. They are presented in approximate order of ascending importance, but are also interdependent.

5.1 The Hunting of the Snark

A forceful critique of the connectionist approach was made by Crick [17]. He argues that connectionist "models" do not correspond sufficiently closely to real neural systems to be regarded as models in the usual sense: "In another context they might reasonably be referred to as existence proofs. As such they have a certain use." [17, p.131]. He attacks connectionist research on methodological and ontological grounds.

His methodological objection is that connectionists employ mathematics as an intellectual prop: he suggests "...that within most modellers a frustrated mathematician is trying to unfold his wings. It is not enough to make something that works. How much better if it can be shown to embody some powerful general principle for handling information, expressible in a deep mathematical form, if only to give an air of intellectual responsibility to an otherwise rather low-brow enterprise." [17, p.132]. A similar criticism has also been made by Lehnert, who argues that "the interdisciplinary appeal of connectionism is not so much a computational appeal, as it is an appeal based on theorem envy. ...connectionism has come to the rescue of a new generation of psychologists who are really closet mathematicians and physicists." [56, p.40].

Crick criticises past work in psychology and linguistics for its preference for simple models which achieve the task in hand in an intelligible manner, an approach rooted in the belief that the brain is intractably complex. He notes that proponents of such an approach are generally not concerned by the absence of criteria of biological or psychological feasibility, and characterises their philosophy thus: "If it describes, in a succinct way, some of the psychological data, what can be wrong with that? Notice, however, that by using such arguments, one could easily make a good case for alchemy or the existence of phlogiston." [17, p.131]. He continues:

> "Why not look inside the brain, both to get new ideas and to test existing ones? The usual answer given by psychologists is that the details of the brain are so hor-

of this paper all three terms will be treated as synonymous, using "connectionism" as an umbrella term. Once again, this paper concentrates only on connectionist models which are also simplifying computational neuroscience models: engineering or purely psychological models are ignored. Examples of all three styles can be found in [50, 42].

It is is not within the scope of this paper to attempt a comprehensive definition of connectionism, nor is it possible to fully examine here the position of connectionism in the context of past practice in cognitive modelling. However, both of these tasks have been performed by Smolensky [56], and the notes that follow in this section are drawn wholly from his comprehensive and influential article. They are presented here as representative of the field, rather than as the view of a small group of Smolensky cohorts. This section presents his definition of connectionist models, and summarises his characterisation of the relationship of connectionism to preceding paradigms of cognitive modelling. The section after this one discusses criticisms of the approach.

> "Connectionist models are large networks of simple parallel computing elements, each of which carries a numerical *activation value* which it computes from the values of neighboring elements in the network, using some simple numerical formula. The network elements, or *units*, influence each other's values through connections that carry a numerical strength, or *weight*. ...if a unit has a positive activation value, its influence on a neighbor's value is positive if its weight to that neighbor is positive, and negative if the weight is negative. In an obvious neural allusion, connections carrying positive weights are called *excitatory* and those carrying negative weights are *inhibitory*.
>
> "In a typical connectionist model, input to the system is provided by imposing activation values on the *input units* of the network; these numerical values represent some encoding, or *representation* of the input. The activation on the input units propagates along the connections until some set of activation values emerges on the *output units*; these activation values encode the output the system has computed from the input. In between the input and output units there may be other units, often called *hidden units*, that participate in representing neither the input nor the output.
>
> "The computation performed by the network in transforming the input pattern of activity to the output pattern depends on the set of connection strengths; these weights are usually regarded as encoding the system's knowledge. ...many connectionist networks *program themselves*, that is, they have autonomous procedures for tuning their weights to eventually perform some specific computation. Such learning procedures often depend on training in which the network is presented with sample input/output pairs from the function it is supposed to compute. In learning networks with hidden units, the network itself "decides" what computations the hidden units will perform; because these units represent neither inputs nor outputs, they are never "told" what their values should be, even during training." [56, p.1; original emphasis]

Smolensky describes connectionism's relation to the two main levels of analysis in previous work: the neural (as in realistic computational neuroscience) and the symbolic (as in nearly all work in cognitive modelling and artificial intelligence (AI) up until the early 1980's). Smolensky posits connectionism as constituting a new level of analysis, intermediate between the symbolic and neural levels. This he refers to as the *subsymbolic* level. The argument for the existence of this new level is involved, and will not be repeated here; see [56].

Smolensky [56, pp.59–62] describes a spectrum of positions on connectionism's relation to the symbolic approach. The spectrum he introduces has as its two extremes *eliminativism*

2. The results may be invalidated by the inadvertent exclusion of important features, because all the cellular details are not yet known.

3. Realistic models incur high computational costs. Progress has been hampered by the lack of available computing power, and many theoretical models and questions make computational demands beyond the state of the art. Until quite recently, only small nervous systems, or fragments of more complex ones, have been feasible as realistic models.

Of these weaknesses, the third will be treated here as a pragmatic issue, and ignored. The first two, particularly the second, will be returned to later in this paper.

Probably the best known realistic models are the Hodgkin-Huxley neuron model [28] and the Hartline and Ratliff model of lateral inhibition in the eye of *Limulus* [48]; many other realistic models are discussed in [39, ch.2–6]. The realistic modelling approach is most appropriate when knowledge of the circuit to be modelled is almost complete down to the biophysical level, and the function of the circuit is already known [38, p.361].

3.2 Simplifying models

The use of "simplifying" models is an approach applied at the network level of nervous systems: the procedure involves starting with a function such as a perceptual ability and designing "simplified neural circuits that can perform the function within the constraints of the state of knowledge." [38, p.361].

Proponents of simplifying models claim that they offer greater conceptual clarity, and that they fulfill a perceived need for models capturing important principles. Their supporters propose [52, p.1300] that they are analogous to aspects of the physics literature such as textbook examples that admit exact solutions, or minimal models (e.g. phase transitions) which reproduce the essential properties of physical systems:

> "The study of simplifying models of the brain can provide a conceptual framework for isolating the basic computational problems and understanding the computational constraints that govern the design of the nervous system." [52, p.1300].
>
> "These models abstract from the complexity of individual neurons and the patterns of connectivity in exchange for analytical tractability." [52, p.1301].

Sejnowski *et al.* state [52, pp.1300–1301] that simplifying models include those currently being investigated under the general headings of connectionist models, parallel distributed processing (PDP) models, and "neural networks"; they remark that some of the models reported on under these headings are used not for brain research, but for neural engineering or for the study of purely psychological processes. It is those connectionist models used for brain research that are discussed here. Note that in this paper I'm not questioning the general principle of studying simplified systems: Galileo's studies of balls rolling down inclined planes are a good example of the usefulness of extremely simple models (see e.g. [58, pp.28–32]). What is under scrutiny in this paper is whether the current *style* of simplifying computational neuroscience models is misguided or not.

4 Connectionism

There has been some debate over the meaning of the terms "parallel distributed processing", "connectionism", and "neural networks". These terms have been loosely applied in the past, to a number of somewhat diverse and underdeveloped theoretical frameworks. For the purposes

assume here that the sole domain of cognitive science is psychologically or biologically plausible models of cognitive processing. That is, I assume that the creation of what Dennett [18] calls "cognitive wheels"[3] is properly only the domain of AI; although, of course, some cognitive wheels may be lurking within the cognitive science canon, yet to be refuted.

3 Computational neuroscience

The notes in this section are taken mainly from [52] and various parts of [34], to which the reader is referred for further information.

The advantages claimed [52, p.3000] for modelling and simulation in computational neuroscience over conventional experimental techniques include:

1. Increased accessibility to the consequences of complicated nonlinear brain systems with many interacting components.
2. The possibility of discovering new phenomena by comparing experimental results to the predictions of simulation, and using these predictions as the basis for the design of new experiments.
3. Facilitating experiments (such as selective lesioning or ablation of particular channels, synapses, neurons or pathways) which would be difficult or even impossible to perform on living tissue.

Practitioners of computational neuroscience employ two classes of brain model: *realistic* models and *simplifying* models;[4] these are discussed in more detail below. Briefly, simplifying models are driven primarily by functional considerations, while realistic models are motivated more by signal measurements and anatomy. In this sense, as Sejnowski *et al.* remark [52, p.1302], simplifying models provide only the most general guidance as to what really happens in the brain, and there is no guarantee that realistic models are not concentrating on aspects of the signals that are irrelevant for information processing. It is also important to realise that the two classes of model represent the extreme points of a continuum: in practice, models are likely to have features of both, and many different types of model will be required to span all levels of the MacGregor-Lewis stratification:

> "...a model of an intermediate level of organisation will necessarily simplify with respect to the structural properties of lower level elements, though it ought to try to incorporate as many of that level's functional properties as actually figure in the higher level's computational tasks." [52, p.1305].

3.1 Realistic Models

A realistic model "...consists of a very large scale simulation that tries to incorporate as much of the cellular detail as is available" [52, p.1300]. The realism of the model, while offering the advantages listed above, also leads to some (potential) weaknesses [52, p.1300]:

1. As more parameters and variables are added to the model to increase its realism, the complexity of the model grows, and so there is an increasing danger of the simulated nervous system being as poorly understood as the real thing.

[3] "... an elegant but unnatural solution to a problem of natural design." [14, p.65]

[4] This terminology is, I assume, not meant to mask the requirement of *any* worthwhile model — that it simplifies whatever system is under study. Perhaps *quite simple* and *extremely simple* would be a more appropriate classification.

there is a sense in which the behaviour of variables at the upper level is explained in terms of variables at the lower level. (2) Models which skip a level are difficult to test and generally low in believability. (3) Models which interrelate variables at a single level are weak in predictive power. (4) Models which relate variables of several strata are most broadly significant."

2 Neurosomething

The recent excitement in the cognitive science and artificial intelligence communities about "neural" issues has been motivated by a number of interests, and consequently the terminology has not been developed in a uniform manner. The first requirement is to distinguish between issues of science and issues of engineering. To this end, two important terms, computational neuroscience and neural engineering, are informally defined below.

- *Computational Neuroscience* is the study of real brains or nervous systems as computational systems, "...in the sense of representing, processing, and storing information." [33, p.3]. Sejnowski and Poggio define it as:

  > "...an approach to understanding the information content of neural signals by modelling the nervous system at many different structural scales, including the biophysical, the circuit, and the systems level." [53]

  Most often, this entails modelling the system under study on a digital computer, but this is neither a necessary nor a sufficient condition for the study to be within computational neuroscience: theoretical analysis using pencil and paper may well suffice [33, p.4]. The term *computational neurobiology* [2, p.186] is here taken to have meaning identical with *computational neuroscience* [*ibid.* p.203].

- The construction of intelligent systems and massively parallel (fine-grain) digital computers inspired by principles found in naturally occurring neural systems has been referred to as *neural computing* [2, p.186] and *neural engineering* [2, p.203]. In this paper only the latter term will be used. The most important feature that distinguishes neural engineering from computational neuroscience is that in neural engineering the emphasis is on building a network of processors which maximises some evaluation metric that is unconcerned with biological reality. That is, neural engineering aims to create artefacts which satisfy constraints based on considerations of (e.g.) processing speed, efficiency, space, cost, reliability, maintainability, etc.: whether the resulting computational system could feasibly be implemented by a real brain or nervous system is irrelevant.

Computational neuroscience and neural engineering are not mutually exclusive endeavours: "...the two subjects have different goals. Nonetheless, they overlap and are mutually stimulating." [2, p.186]. Nevertheless, the science/engineering distinction is an important one:

> "Engineering is often based on science, but its aim is different. A successful piece of engineering is a machine which does something useful. Understanding the brain, on the other hand, is a scientific problem." [17, p.132].

This paper is concerned only with science; not with engineering. In particular, it is concerned with the relationship of computational neuroscience to cognitive science. It is not concerned with any attempts to create 'intelligent' artefacts whose primary design constraints ignore biological or psychological reality. Such artefacts are part of the domain of artificial intelligence (AI), and for the purposes of this paper cognitive science and AI will be considered as distinct disciplines. I

the fruits of computational neuroethology simulations are more "hard" objective measurements than "soft" subjective ones. At a metatheoretical level, it is argued that the computational network simulation of cognitive processing should pay much more attention to the evolutionary history of those faculties it wishes to replicate. In particular, a conclusion of this paper is that the study of linguistic processes using network models is wildly premature.

As the reader will probably already have detected, this paper is intentionally polemic. It is aimed at an interdisciplinary audience, and the author is no polymath: some of the informal definitions of terms may be found wanting, and there might be some holes in the argument. For that reason, this paper is offered as a provisional manifesto in the hope that it provokes some interesting discussion. The argument is based on previous work by a number of authors. Because of its disputatious nature, there are more direct quotes in this paper than is common. Where the quotes are not felicitous, they are used because to paraphrase is to invite charges of distortion. Nevertheless, there is no denying that this is a *selective* review of the literature.

The paper opens with a discussion of computational neuroscience, distinguishing it from neural engineering, and identifying the two classes of model: realistic and simplifying. Following this, the connectionist paradigm is briefly summarised. Next, criticisms of connectionism are discussed, with particular attention to the argument that connectionist models have no semantic grounding without a sensorimotor system. Following this, a remedy to such objections is proposed: the adoption of the computational neuroethology approach. Computational neuroethology is defined, and a specific approach to providing a behavioural linkage is discussed. This approach has some important implications for future research, the most significant of which is that it encourages an evolutionary approach to understanding intelligence; an argument rooted in previous work by, among others, Hoyle and Brooks.

1.1 The MacGregor-Lewis Stratification

The discussion in this paper employs the MacGregor-Lewis stratification of neuroscience. This is a simple taxonomy of levels of analysis.[2] The levels of the stratification [39, p.8] are listed below. Computational neuroscience theory aims mainly at linking levels C and D.

E Behaviour, experience, consciousness.
(Direct observations.)

D Theoretical inferred constructs.
(Superego, id, drives, cognitive maps, cell assemblies, statistical configurations.)

C Electrical signals.
(Field potentials, spike trains, generator potentials.)

B Membrane conductance modulations.
(Synaptic conductances, active dendritic conductances, action potential conductances.)

A Molecular and chemical processes.
(Chemical synaptic transmission, gating processes.)

As MacGregor [39, p.8] notes:

> "Several principles concerning modelling are suggested by this stratification of variables. (1) Models which relate variables of adjacent strata are the most powerful;

[2] Use of this stratification should not be taken as absolute endorsement of it. It is merely a convenient taxonomy for the purposes of this paper: it could be improved. In particular, level D seems so broad as to be unwieldy. I think that level D could usefully be divided into: D1 (Biological inferred constructs); and D2 (Psychological inferred constructs).

Computational Neuroethology: A Provisional Manifesto

D. T. Cliff

School of Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QN
England, U.K.

E-mail(JANET):davec@uk.ac.sussex.cogs

May 1990

Abstract

This paper questions approaches to computational modelling of neural mechanisms underlying behaviour. It examines "simplifying" (connectionist) models used in computational neuroscience and concludes that, unless embedded within a sensorimotor system, they are meaningless. The implication is that future models should be situated within closed-environment simulation systems: output of the simulated nervous system is then expressed as observable behaviour. This approach is referred to as "computational neuroethology". Computational neuroethology offers a firmer grounding for the semantics of the model, eliminating subjectivity from the result-interpretation process. A number of more fundamental implications of the approach are also discussed, chief of which is that insect cognition should be studied in preference to mammalian cognition.

1 Introduction

This paper is concerned with approaches to computational modelling of the neural mechanisms underlying behaviour. It examines the relationship between computational neuroscience (e.g. [52, 34]) and that style of modelling popularly referred to as "connectionism", "parallel distributed processing" (PDP), or "neural networks"[1] which has recently been subject to renewed attention in the fields of cognitive science and artificial intelligence (see e.g. [50, 42]).

Connectionist models are characterised by their simplified nature and concomitant inattention to biological data, and it is argued here that such "simplifying" computational neuroscience has serious inadequacies. A different approach is suggested which pays far more attention to the sensorimotor system and hence to behavioural interactions with the external environment. This approach involves computational modelling of the neural mechanisms underlying behaviour, in a manner akin to that used in connectionism. Meaning is supplied to the models by embedding them in simulated environments which close the external feedback loop from motor output to sensory input. That is, they supply sensory feedback without human intervention. Such an analysis of behaviour as a product of neural activity is properly the domain of the field of neuroethology, and the new approach is therefore referred to as "computational neuroethology".

The advantage of computational neuroethology is that the semantics of the network are well grounded, and thus results are generated by observation rather than by interpretation. That is,

[1] In this paper, these three terms will be treated as synonymous, and referred to collectively as "connectionist" models. This form of connectionism differs from the style of neuroethology research sometimes referred to as connectionism (e.g. [21, p.337]).

Computational Neuroethology: A Provisional Manifesto

D. T. Cliff

May 1990

Cognitive Science Research Paper

Serial No. CSRP 162

The University of Sussex
School of Cognitive and Computing Sciences
Falmer
BRIGHTON BN1 9QN
England, U.K.