Smoke Without Fire:
What do virtual experiments in Cognitive Science really tell us?

Peter R. Krebs (peterk@cse.unsw.edu.au)
Cognitive Science Program
History & Philosophy of Science
The University of New South Wales
Sydney, NSW 2052, Australia

Abstract

Many activities in Cognitive Science involve complex computer models and simulations of both theoretical and real entities. Artificial Intelligence and the study of artificial neural nets in particular, are seen as major contributors in the quest for understanding the human mind. Computational models serve as objects of experimentation, and results from these *virtual* experiments are tacitly included in the framework of empirical science. Simulations of cognitive functions, like learning to speak, or discovering syntactical structures in language, are the basis for many claims about human capacities in language acquisition. This raises the question whether results obtained from experiments that are essentially performed on data structures are equivalent to results from "real" experiments. This paper examines some design methodologies for models of cognitive functions using artificial neural nets. The process of conducting the cognitive simulations is largely a projection of theories, or even unsubstantiated conjectures, onto simulated neural structures and an interpretation of the experimental results in terms of the human brain. The problem with this process is that results from virtual experiments are taken to refer unambiguously to the human brain; and the more the language of human cognitive function is deployed in both theory construction and (virtual) experimental interpretation, the more convincing the impression. Additionally, the complexity of the methodologies, principles, and visualization techniques, in the implementation of the computational model, masks the lack of actual similarities between model and real world phenomena. Some computational models involving artificial neural nets have had some success, even commercially, but there are indications that the results from virtual experiments have little value in explaining cognitive functions. The problem seems to be in relating computational, or mathematical, entities to real world objects, like neurons and brains. I argue that the role of Artificial Intelligence as a contributor to the knowledge base of Cognitive Science is diminished as a consequence.

Introduction

Models and simulations have been described as mock-ups, analogies, simplifications, or metaphors, and sometimes the term *simulation* carries connotations of pretense, or even deceit. However, in the world of science and technology, these connotations have largely faded, and the use of computational models and simulations has become common practice. Fox Keller (2003) remarks that during the 1940s

> the valence of the term [simulation] changes decisively: now productive rather than merely deceptive, and, in particular, designating a technique for the promotion of scientific understanding (Fox Keller, 2003, p198).

While models are generally accepted as tools in the empirical sciences, their epistemological status is still to be determined. Ziman (2000) points out that the notion of a model defies formal definition like other metascientific concepts. Without a proper definition of what is understood by a model, it is also difficult to place models into a scientific framework. Models have served as the basis for major shifts in theories in the physical sciences. Bohr's model of atoms, for example, changed the way in which chemists could predict the properties of substances. Models and simulations can also be in the form of some apparatus, like the model of an aeroplane in a wind tunnel, or the ball and stick models of molecules. Computer models and simulations (CMS)[1] are a progression from computational utility. In the early days of modern computing, the range of problems that could be subjected to quantitative analysis had been radically extended (Fox Keller, 2003). Calculations, that had been too complex and tedious to deal with numerically, became trivial in a very short period of time[2]. CMS possess properties that extend the utility of previous models in terms of speed and numerical accuracy. They allow for convenient 'what if' experiments where the behavior of a mathematical model can be examined over a range of changing parameters. On this basis, CMS seem to surpass theoretical models, such as Bohr's atom model, or physical models that are made of real material. However, there are particular problems in using CMS as objects of

[1] I will use the terms computer model and computer simulation interchangeably. In the context of this paper models and simulations are mathematical constructs that have been instantiated as executable programs. A *simulation* is a model that has been designed to illustrate its dynamics, but a clear distinction is neither possible nor necessary in the context of this paper.

[2] Monstrous analog tide calculating machines, high precision developments of Lord Kelvin's tide predictors, were operating until the mid 1960s. Williams (1997) notes that the machine, constructed and operated by the U.S. Coast and Geodetic Survey, could calculate the height of the tides to the nearest 0.1 ft for each minute of the year for a location in a few minutes. The magnitude of calculations involved to establish a tidal forecast can be gaged by the fact that in a modern computer the cosine sub-routine is called about 20 million times to predict the tides for a single year for a single location (Williams, 1997). The tide predictors were essentially mechanical models of the cyclic movement of heavenly bodies. Nowadays, these calculations can be solved numerically with higher precision in a few seconds.

scientific experimentation and also as tools to support claims about theories.

Experimentation

CMS in engineering disciplines or the physical sciences are dealing with real world entities and phenomena. Some mapping from the real entities to the modeled entities usually exists and such 'realistic models' are approximate representations of the real world. How approximate do the models have to be for *scientific* experimentation? For Hacking, experimentation is not merely about the observation of phenomena and the subsequent inferences about the underlying theories, but is about observing and *interfering* with the objects in question. The ability to manipulate objects is an essential part of the experiment, which is "to create, produce, refine, and stabilize phenomena" (Hacking, 1983, 230). The close connection between experiment and some real world entities is also a key requirement in the definition offered by Harré, who says that

> [a]n experiment is the manipulation of [an] apparatus, which is an arrangement of material stuff integrated into the material world in a number of different ways (Harré, 2003, p19).

The kinds of experiments that fit the criteria, which relate to the discussions by Harré and Hacking, are the activities we often associate with what happens in the laboratory. These are the kinds of experiments we know about from our high school days. However, it has become obvious that the vast majority of experiments are different from this stereotypic view (Morgan, 2003). There are no materials that could possibly be manipulated in experiments with CMS. The material, the apparatus and the process of interference are all replaced by data structures and computational processes[3]. The nature of the entities and the phenomena that are the points of interest in the field of Cognitive Science dictates that CMS are often the only way to do any experimentation at all. The experiment is moved into the realm of the *virtual* for convenience or necessity.

Assumptions and Methods

In the current Theory of Mind in Cognitive Science it is a fundamental assumption that cognitive functions are computational processes, or at least computable processes, performed by brains. This hypothesis is attractive, because, as Sterelny puts it,

> [i]t is good research strategy to try to model our information processing on something we already know a bit about. And we do know a good deal about computation, both from the theory of formalised systems of reasoning and from the actual implementation of some of those systems on real machines (Sterelny, 1989, p74).

It seems obvious that CMS would be the ideal approach to provide the clues and explanations for the Theory of Mind, as computation is the essence of what happens in brains and in CMS.

Viewing the brain as a *black box* that computes is only one level of description. Another takes into account that brains are amongst other things composed of about 10^{10} neurons that are interconnected by about 10^{14} synapses. The neuron has been determined as the smallest building block in terms of computational power[4]. CMS can be based on either assumption, resulting in models that treat either the brain or the neuron as a *black box*. In Artificial Intelligence these levels of description correspond in many ways to the symbolic and connectionist paradigms. The symbolic approach is primarily concerned with *what* goes on in the brain, while the connectionists are interested in *how* things happen in the brain. The connectionist approach adds a kind of neural plausibility, because these modeling techniques are methodologically comparable with what neurons do in brains (Schultz, 2003).

At this point, it is important to note that the description of the functionality of a neuronal model in AI, is already far removed from the actual observable functions of a neuron, even if we restrict the discussion to the most basic *input*/*output* functionality of the biological neuron[5]. The theory about the outward computational functionality of the simplified neuron does not take this behavior into account. The dissimilarities initially concerned hardware based models, but shifting these models into the 'virtual' does not reduce, or eliminate, these discrepancies. For example, the *rate* of switching is of importance in biological neurons[6] , but it has no significance in the neural models used commonly in artificial neural nets (ANN). Also, the number of connections between neurons is much larger than most models take into account. Some neurons may have 10^6, or 10^7, synapses, whereas the number of connections in models is usually restricted to less than 10^2. Modeling techniques at neural level are only loosely aligned with the real biology and the real observable behavior of neurons.

Construction

The earliest functional models of neurons by McCulloch and Pitts (1943) were simple switching devices and the modeled behavior was strictly according to the rules of elementary logic. The discreet components, which only operated in

[3] The only material part of the experiment is the computer hardware. With the proliferation of the personal computer it turns out that the vast majority of CMS are implemented on the same platform, i.e., more or less identical hardware.

[4] Penrose (1990) and others propose that some form of quantum computing is performed with the neuron.

[5] For example, real neurons change their output behavior in response to the firing *rate* of neighboring neurons. The rate at which information is presented to the inputs (synapses) has a *real* effect on the eventuality and timing of the neuron's output (spiking).

[6] The firing rate changes the behavior dynamically, so that a neuron is more sensitive after a burst of activity.

switching mode, i.e. *on*/*off*, were designed to implement the basic logic functions *AND*, *OR*, and *NOT*. These early models had little in common with real neurons and the experiments did not yield much in terms of progress toward an artificial intelligence. With hindsight, Copeland points out that

> [h]alf a century later on, it is clear that people were putting two and two together to make five. Even at that time there was a certain amount of nagging evidence regarding the dissimilarities between neurons and computing hardware (Copeland, 1993, p185).

Hebb's discovery, that some neural connections are modified over time by patterns of excitations (Hebb, 1949), led to the development of the *perceptron*, which represented one of the first learning networks (Rosenblatt, 1958, 1962). A decade later, the connectionist program came almost to a halt after the "remorseless analysis of what perceptrons cannot do" (McLeod et al., 1998, 323). This analysis by Minsky and Papert (1988) showed that perceptrons can *only* solve tasks that are *linearly separable*[7]. Neural network models that employ primary elements of the *all-or-nothing* type without sophisticated transition functions offer little, if anything, beyond digital circuitry. Since then, artificial multi-layered networks comprising non-linear neurons have been shown to be much more powerful than Minsky and Papert were willing to admit. Churchland writes that

> [...] a nonlinear response profile brings the entire range of possible nonlinear transformations within reach of three-layer networks [...] Now there are *no* transformations beyond the computational power of a large enough and suitably weighted network (Churchland, 1990, p206).

and Elman et al. say that

> [f]or some ranges of inputs [...] these units exhibit an all or nothing response (i.e., they output 0.0 or 1.0). This sort of response lets the units act in a categorical, rule-like manner. For other ranges of inputs, however, [...] the nodes are very sensitive and have a more graded response. In such cases, the nodes are able to make subtle distinctions and even categorize along dimensions which may be continuous in nature. *The nonlinear response of such units lies at the heart of much of the behavior which makes networks interesting* (Elman et al., 1998, p53).

As Elman et al. point out, the nonlinear and continuous transition functions of more 'nature-like' neural models seem to offer much more. The question is, of course, how much more? If we examine the mathematics of the models[8] it becomes apparent that there is only some rudimentary, if not superficial, functional similarity on offer[9]. This kind of simplicity remains in the assemblies of these units, ANNs. Feed forward and recurrent networks implement *functions* from the inputs to the outputs - these networks perform *linear* or *non-linear regression*. The connections between the nodes in the hidden layer contain information about the mappings of these functions, and these mappings become the source of 'insights' about what goes on in the ANNs. It is possible, with appropriate statistical methods, to map this information (activation patterns) onto locations in *n*-space to make new claims about what the models achieved.

Interpreting Models

The assumption that symbols and their semantic contents are distributed throughout the network is part of the connectionist doctrine. However the interpretation of experimental results in the context of neural nets is not possible without the use of symbols at some level of description. Hoffmann points out that

> [...] in more complex systems, the use of symbols for describing abstractions of the functionality at the lowest level in inevitable [...] any description of a sufficiently complex system needs layers of abstraction. Thus, even if a non-symbolic approach uses tokens at its base level which cannot be reasonably interpreted, there still needs to be a more abstract level of description (Hoffmann, 1998, p257).

For a meaningful interpretation of the network and its dynamics, it is necessary to convey content and meaning in terms of non-distributed, or localised, symbols. Elman et al. suggest that

> localist representations [...] provide a straightforward mechanism for capturing the possibility that a system may be able to simultaneously entertain multiple propositions, each with different strength, and that the process of resolving uncertainty may be thought of as a constraint satisfaction problem in which different pieces of information interact (Elman et al., 1998, p90).

Localized representations and distributed representations cannot be used together in a single representational system without confusing the semantic meaning of representations. ANNs are described as having distinct and discrete inputs and outputs, each labeled with distinct and discrete meaning. These localized representations are no longer available once the focus shifts on to hidden nodes within the network, and the 'representations' are now described in terms of weights,

[7] The relatively simple *XOR*-function cannot be handled by a single-layer network of perceptrons.

[8] Refer to Haykin (1999), Russel and Norvig (1995), Lytton (2002), Hoffmann (1998), and many others.

[9] Neural models in the field of Computational Neuroscience are much more aligned with the biology and chemistry of neurons, but they are not the kind of model employed in AI.

or synaptic strengths, between individual units. The mistake, I believe, is to bring the top-down psychological model and the bottom-up neural environment together and to treat the result as a coherent and meaningful combination.

Elman (1990), for example, presented bit patterns to a recurrent network, where each individual bit represented a particular word in the human language. The patterns themselves were presented in sequences forming two and three word sentences that had been generated according to a set of fixed templates. A cluster analysis of the hidden nodes revealed that the trained network exhibits similar activation patterns for inputs (words) according to their relative position in the sequence (sentence). The analysis of these activation patterns allowed for the classification of inputs (words) into categories like *nouns* or *verbs*. It is important here to understand that these results are not furnished by the ANN, instead they are *interpretations* of internal structures at a higher level. The actual role of the ANN is that of a *predictor*, where the network attempts to predict the next word following the current input[10]. If the ANN is meant to be a model of what happens at the neural level, then the question arises, what mechanism is responsible for the equivalent analysis of activation patterns in the brain. We will have to assume another neural circuit to do an analysis of the hidden nodes. This new network could categorize words into *verbs* and *nouns*, but then we need another circuit to categorize words into *humans*, *non-humans*, *inanimates*, or *edibles*, and another to categorize words into *mono-syllabic* and *multi-syllabic*. In fact, we will need an infinite number of neural circuits just for the analysis of word categories.

Churchland (1998) describes a recurrent network that could model more challenging cognitive functions. He suggests that a recurrent network may have an appropriate architecture for learning and simulating moral virtues. He considers, given the "*examples* of perceptual or motor categories at issue", that a network would be able to map concepts like *cheating*, *tormenting*, *lying*, or *self-sacrifice* within a *n*-space of classes containing dimensions of *morally significant*, *morally bad*, or *morally praiseworthy* actions. Churchland says that

> [t]his high-dimensional similarity space [...] displays a structured family of categorical "hot spots" or "prototype position", to which actual sensory inputs are assimilated with varying degree of closeness (Churchland, 1998, p83).

I believe that this approach toward a calculus of moral virtues is flawed for two reasons. First there is the question of what kinds of "actual sensory inputs" could be available to train a network in moral virtues, or to condition a brain in moral virtues. The second problem is whether moral viewpoints can be synthesized from a possibly large number of discrete constituents. For this approach to work, it would have to be possible to define a moral action as the function over a set of discrete inputs. However, a morally bad action like *stealing an item of clothing* is not simply the result of *poverty = true* and *night time = true* or *low temperature = true* and *coat available = true*, and so on. If it were so, then our lives would need to be expressed in terms such that a set of mathematical functions could determine our next action, a proposition that has profound philosophical consequences.

Whether one subscribes to Chomsky's notion of a universal grammar or not, relationships between syntax and semantics do exist in natural language. It is these relationships that were explored by Elman (1990). It should be clear that a grammar has far less rules than what makes up the moral fabric of a human being. Unlike the formalisms that are evident in natural language, there are no similar formalisms available for the analysis of moral virtues by means of non-linear regression.

A more interesting problem lies in the interpretation of representations that are within the network. First there is the question of *locating* suitable representations that could carry any semantics, given that the representations are distributed in the network. Rosenblatt explained that

> [i]t is significant that the individual elements, or cells, of a nerve network have never been demonstrated to possess any specifically psychological functions, such as "memory", "awareness", or "intelligence". Such properties, therefore, presumably reside in the organization and functioning of the network as a whole, rather than in its elementary parts (Rosenblatt, 1962, p9).

However in CMS, the input nodes and output nodes are treated as localized representations (symbols). Individual model neurons *do* have semantics bestowed upon them by Elman (1990) and Churchland (1998), who map *meaningful* words and moral concepts to the inputs and outputs of their networks. Treating the activation patterns of the hidden units as a resource for categorical "hot spots", to use Churchland's term, is an even more contentious exercise. The relationships and patterns in the input datasets and training datasets become embedded in the structure of the network during training[11]. The internal representations, which are "snapshots of the internal states during the course of processing sequential inputs" (Elman, 1990), are extracted by means of cluster analysis of the hidden layer in the ANN. Who really *does* the analysis and interpretation of the distributed representations? The experimenter performs these tasks using a new tool, i.e. cluster analysis, - the network has no part in this. Moreover, an appropriate analysis, performed on the training data, could yield the same information. The networks merely compute functions, and the activities of the networks do not add any *additional* information. Despite all the complexities of the mathematics of ANNs, the functions that are performed are relatively trivial.

[10] The actual word, which follows the input in the training set, is used as the target to determine the error for back propagation during the training phase.

[11] The patterns and relationships in these datasets can either be carefully designed or might be an unwanted by-product.

Conclusions

A model is a simplification of its real world counterpart, but models must maintain some plausible connection to real world objects or real world phenomena. Green (2001) suggested, that some of the 'apparent' successes of connectionist modeling may well be based on a rather vague concept of what *is* actually modeled. The simple *functional* neurons that are employed in AI, resemble only loosely actual biological neurons, and ANNs exhibit only superficial commonalities with brain structures. The building blocks and tools used in the connectionist paradigm of AI offer some plausibility for the bottom-up approach nevertheless. While many models share by design the connectionist *architecture*, the processes and functions under investigation seem quite different. The investigations about language, moral virtues, and many other topics, belong to the top-down approach where localized representations are used to convey the semantic contents of sensory and conceptual entities. Merging the two opposing paradigms within models is not without problems. We can easily assign meaning to localized representations, and we can manipulate representations without loss of semantics, provided we maintain appropriate syntactic rules. The processes break down when localized representations are 'manufactured' by assigning them to concepts seemingly emerging from ANNs. The danger is that statistical artifacts are presented as novel phenomena of the model. However, there are no novel phenomena emerging and there is nothing intelligent happening within - there is no fire, not even a spark.

Acknowledgment

I would like to thank Anthony Corones for his encouragement, valued comments and suggestions on earlier drafts of this paper.

References

Branquinho, J. (2001). *The foundations of cognitive science*. Oxford: Oxford UP.

Churchland, P. M. (1990). *Cognitive Activity in Artificial Neural Networks*. In Cummins and Delarosa Cummins (2000).

Churchland, P. M. (1998). *Toward a Cognitive Neurobiology of the Moral Virtues*. In Branquinho (2001).

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Malden: Blackwell.

Cummins, R. and Delarosa Cummins, D. (2000). *Minds, Brains, and Computers: The Foundations of Cognitive Science*. Malden: Blackwell.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L., Bates, E. A., Karmilof-Smith, A., Parisi, D., and Plunkett, K. (1998). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, Massachusetts: MIT Press.

Fox Keller, E. (2003). *Models, Simulation, and "Computer Experiments"*. In Radder (2003).

Green, C. D. (2001). Scientific models, connectionist networks, and cognitive science. *Theory & Psychology*, 11(1):97–117.

Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambrigde UP.

Harré, R. (2003). *The Materiality of Instruments in a Metaphysics for Experiments*. In Radder (2003).

Haykin, S. (1999). *Neural Nets: A Comprehensive Foundation*. Upper Saddle River, New Jersey: Prentice-Hall.

Hebb, D. O. (1949). *The Organization of Behavior*, chapter 19. In Cummins and Delarosa Cummins (2000).

Hoffmann, A. (1998). *Paradigms of Artificial Intelligence*. Springer.

Lytton, W. W. (2002). *From Computer to Brain: Foundations of Computational Neuroscience*. Springer.

McCulloch, P. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McLeod, P., Plunkett, K., and Rolls, E. T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford UP.

Minsky, M. and Papert, S. (1988). *Perceptrons : An Introduction to Computational Geometry (2nd Edition)*. Cambridge, Massachusetts: MIT Press.

Morgan, M. S. (2003). *Experiments without Material Intervention*. In Radder (2003).

Penrose, R. (1990). *The Emperor's New Mind*. London: Vintage.

Radder, H. (2003). *The Philosophy of Scientific Experimentation*. Pittsburgh: University of Pittsburgh Press.

Rosenblatt, F. (1958). The perceptron: A probabalistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Washington: Spartan.

Russel, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Prentice Hall.

Schultz, T. R. (2003). *Computational Developmental Psychology*. MIT Press.

Slezak, P. and Albury, W. R. (1989). *Computers, Brains and Minds: Essays in Cognitive Science*. Dordrecht: Kluwer.

Sterelny, K. (1989). *Computational Functional Psychology: Problems and Prospects*. In Slezak and Albury (1989).

Williams, M. R. (1997). *A History of Computing Technology*. Los Alamitos: IEEE Press.

Ziman, J. (2000). *Real Science: What it is, and what it means*. Cambridge: Cambridge UP.