Modelling for its own sake

A new calculation of a neural network model, while mathematically interesting, seems too close for comfort to the domain of modelling for the fun of it.

SIMPLE models of complicated systems have several functions, of which the chief is to give ordinary mortals a comprehensible picture of a phenomenon. Thus the old schoolboy calculations of the bending of a beam under a load are a representation of how a wooden plank may bridge a stream without leading wise people to expect the calculations to apply to real wooden boards, with their fibrous structure. Sometimes, simple models point to general properties that apply even to the complicated systems they represent; Bohr's atomic model gives a good account of how energy levels crowd together near the ionization limit even though its account of the splitting of energy levels in magnetic fields is incorrect. But model-building can get out of hand, as with the fruitless refinement last century of the luminiferous aether to accommodate electromagnetic radiation.

Is there a danger that the same may now be happening with neural network models of the brain? This sour question is provoked by an intricate calculation of a particular neural network by Ronny Meir and Eytan Domany of the Weizmann Institute (*Phys. Rev. Lett.* **59**, 359; 1987). The answer is that there is indeed a danger that neural networks will become a field of study in their own right, unrelated to the phenomena they model, but wringing useful notions from them will increasingly require discrimination.

Neural network models are mostly elaborations of a crisp formulation by J.J. Hopfield (*Proc. natn. Acad. Sci. U.S.A.* **79**, 2554; 1982). An earlier article on this topic (*Nature* **325**, 11; 1987) should have made Hopfield's contribution clear. The idea is to represent neurons by two-valued entities — neurons may be either ON or OFF — and the state of each may be influenced by the states of all others. If the neurons modelled are those of the sensory cortex, the state of the system is, at any time, the brain's impression of the outside world. Hopfield's achievement was to show how, by a suitable choice of the parameters describing the mutual influence of the neurons, to embed "memories" in this system. The embedded memory states are orthogonal to each other in the sense that arbitrary impressions of the external world which overlap with, or evoke, predominantly one of the embedded states will not also evoke another. The embedded memory states are a little like the eigenstates of a quantum system, and thus a basis for the analysis of arbitrary states.

This model has been extraordinarily stimulating. It offers an explanation of the distributed character of memory, as when people whose brains are damaged by accident or surgery find a general impairment, not a selective loss, of memory. Similarly, it is a model for the process of learning, represented as that of fixing the mutual interactions of neurons to define particular ground states, and where feedback of various kinds may be especially important (*Nature* **328**, 107; 1987).

All kinds of intriguing speculations also suggest themselves. Is lateral thinking a measure of the overlap between supposedly orthogonal ground states? Or a dynamic phenomenon in which the processing of sensory information evokes a sequence of partially relevant images? If the sensory cortex is continuously responsible for processing sensory information, what structures in the brain monitor those events and tell which memories are instantaneously evoked? Can one remember moving pictures (apparently yes, see *Nature* **325**, 11; 1987)? Will the model work for language-learning and, if so, does the common structure of human language for which Chomsky has argued say something about the permisssible eigenstates of that part of the brain? All these are coffee-table questions now made a little more tangible. At this rate, the question "What is consciousness?" will again become respectable.

So it is natural that much energy has been spent on elaborations of the neural network model. Meir and Domany have worked with their own layered neural network in which they simulate the columnar structure of the cortex with a sequence of layers of model neurons, each of which is influenced by all cells of the preceding layer. The model provides a means of transforming an image of the external world at the lowest layer into other representations at succeeding layers. As there is no feedback from upper to lower layers, it is also a model of how a single neural sheet might transform an image of the outside world with the passage of time, in which case the parameters specifying the strength of the interconnections become the dynamic rules for the evolution of the system. On the first layer, the representation of the eigenstates of the system, helpfully called "key patterns", is externally dictated, but their representation in other layers is entirely independent. The object of the exercise is to calculate the likelihood that an external stimulus more or less congruent with a key pattern on the first layer is itself transformed so as to be congruent with the transformed image of that same key pattern on succeeding layers.

With certain assumptions, in particular that the number of neurons in each layer tends to infinity, the conclusion is that this is a high-fidelity system provided that the number of stored key patterns is not too great a fraction (0.269) of the number of cells in each of the layers. But if this critical fraction is exceeded, correlation vanishes between the successively transformed versions of the input stimulus and the stored versions of the key patterns.

Does this help those who seek a clearer picture of how the brain functions? That a system of layered neurons can be calculated exactly is something to be pleased about; part of the price paid for this exactness is that the representations of key patterns in successive layers are supposed independent of each other, which is probably not the case in the real columnar cortex. Even so, the same mathematical tricks will no doubt be useful in the calculation of later network models. That the number of key patterns that can be stored in such a system must be limited if memory is to function well is interesting, although that was evident in Hopfield's work five years ago.

The sharp phase-like transition, as the ratio of stored memories to active neurons in each layer increases above the critical fraction, from a condition in which memories are faithful to one in which memory completely fails, may suggest to some the more or less sudden onset of memory impairment in conditions such as Alzheimer's disease; declining numbers of active neurons allow the ratio to exceed some critical value. On that view, the underlying trouble is just as much the excess baggage of stored key patterns as the decline of the number of active neurons. But that, of course, is unbridled speculation, of which Meir and Domany's paper is innocent. That is why it seems uncomfortably near the borderline of modelling for its own sake. **John Maddox**