A Neurobiological Perspective on Building Intelligent Devices

What is "intelligence", and what problems might arise in building an intelligent device? Are human brains, with their greatly expanded neocortex, the only currently existing intelligent devices? Apart from some unconvincing computer programs, the only known intelligent devices seem to be animals, particularly birds and, especially, mammals. However, there is at least one other clear example of natural "intelligence": all living organisms, notoriously, *appear* to be intelligently designed, even though this appearance is achieved by selective amplification of molecular accidents. This form of natural intelligence (i.e. the Darwinian "algorithm" comprised of iterative replication/mutation/transcription/translation/selection steps) is the only other successful exemplar of "intelligence" and provides a thread through the neocortical labyrinth.

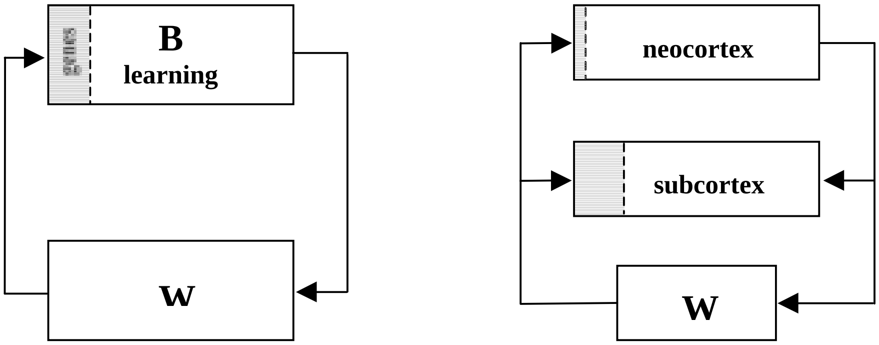

Fig. 1. Left: Interaction between an animal's brain (B) and its world (W). The brain's input-output relation reflects its synaptic weights, which depend on the history of ancestors (gray zone, "genes") and, especially in complex animals, on the history of the animal itself ("learning"). Right: the 2 main components of a mammal's brain: subcortical structures (which learn pairwise correlations) and the neocortex (which learns higher order correlations). The neocortex relies particularly heavily on learning, and provides corrections to subcortical computations.

What is going on inside the skull (Fig 1)? There are 2 basic processes: a rapid (millisecond) "integration" step, in which synaptically weighted voltages are collected over the surface of a neuron, combined (possibly in a nonlinear way), and sent, via more synapses, to other neurons, and a slower "learning" process, which uses the rapid signals to modify the weights such that performance improves. Learning is done by adjusting the strength of *individual* synapses according to the voltage across the synapse (Hebb's Rule). The power of the learned world model will reflect the extent to which the synapses can *individually* be set (much as the power of a digital computer reflects the number of transistors and memory locations that can be individually, and sufficiently rapidly, controlled. Intelligence boils down to numbers: the combinatorial potential vastness of the world should be matched by a corresponding potential combinatorial vastness of the brain that models it, together with precise rules (such as Hebb's) for selecting useful combinations. Integration requires voltage spread but accurate learning requires chemical localization; the incompatibility of these requirements limits intelligence.

The Neocortex: Looking Inside the Box

An enormous amount has been learned about the neocortex. First, it seems to have a similar microstructure from monotremes to Mozart, in different animals and different parts of the same animal. The neocortex has, characteristically, 6 layers. Neocortical input arrives, from a central and mysterious lump of neurons called the thalamus ("layer 0"), in layer 4. The set of input firings, filtered through the 0/4 synapses, initializes a representation which then rapidly evolves as the environment changes and as inhibition and recurrent excitation kick in. This recurrent process can be thought of as providing a statistically optimal estimate of what the initial pattern would have been if there were no noise in the neural circuitry[1]. Thus the core computation is, as originally surmised by Hubel and Wiesel in their pioneering Nobel-winning work on the visual cortex[2], a feedforward input vector-weight matrix –output vector computation that provides an explicit initial representation of the world, which is a linear transformation of an already efficient but less explicit representation furnished by the thalamus.

The thalamic representation is optimal in terms of second order (pairwise) statistics only, while the neocortex takes into account residual, higher-order dependencies. The thalamic representation is merely a copy of the retinal "whitened" or decorrelated representation that is also sent to more primitive brain areas such as colliculus, where immediate actions, based on learned second-order statistics plus inherited knowledge, are initiated. The cortex is an "add-on" device that provides slower, higher-order corrections[3].

Hubel and Wiesel originally suggested that the layer 4 "simple" cells are tuned to local orientation because oriented lines and edges are particularly rich in the natural world (and therefore provide a natural "code"); this insight has been made more quantitative with tools from information theory, statistics etc. Indeed, statistically optimal representation of natural scenes, based on the idea that the mutual information between the scenes and their neural representations should be maximized, leads directly to local orientation filters (Independent Component Analysis: ICA[4]). This strategy exploits higher-order redundancies to generate optimal codes. Such a code is "generative" in that it attempts to model the transformation, in the real world, that leads from "objects" and other underlying "causes" to sensory data (patterns of light on the retina etc). In ICA, the generative model is linear, but in principle nonlinear processes can also be modeled. This seems to happen in the transformation from simple cells (orientation-tuned and position-sensitive) to complex cells (typically found in layers 2 and 3) which are orientation-tuned but locally position-insensitive. Representations in later layers also incorporate temporal correlations (leading, for example, to direction-tuning). This framework is an appealing, though incomplete, "candidate" for the elusive laurel of "canonical microcircuit". (The canonical microcircuit concept, that there is a core information-processing strategy throughout the mammalian cortex, is controversial, but without it prospects for hardware emulation seem hopeless, since there would be nothing to emulate).

The idea that neocortex learns high order statistics, thereby capturing aspects of the structure of the world, and then reprocesses representations using further nonlinearities fits in well with recent evidence that cortical outputs are rerouted back to cortex via the thalamus[5]. This suggests that the neocortex progressively develops a sophisticated world model by hierarchical reapplication of a standard algorithm. This sophisticated model then corrects simpler subcortical processing to generate appropriate behavior.

Accurate model-learning requires specific Hebbian synapses

The brain is a nanoelectronic computing device where information is stored in elementary locations called synapses. Synapses are micron-size units with complex read/write functionalities, but the information is stored as numbers of transmitter-sensitive ion channels, at resolutions ranging from 1 to several dozen bits per synapse. As in dry computers, information is read and written using voltage pulses. The "read" voltage is provided by the arrival of a presynaptic "spike", which releases transmitter, which generates a postsynaptic response proportional to the number of transmitter-sensitive ion channels. The summed postsynaptic responses may trigger a delayed "write" pulse (in the form of a backpropagating dendritic spike, which initiates a small synaptic strength increment at synapses marked by the arrival of a "read" spike in the previous 10 msec interval.) The process of incrementing those synapses whose read-out contributed to “write” spikes is known as a “Hebb Rule”. The overall effect of the Hebb rule is that neurons, and hence brains, steadily improve their predictive abilities (output spikes become better correlated with input spikes, much in the way that organisms steadily improve their replicative ability). But, just as in a dry computer, a fundamental limit is set by the precision with which information can be cheaply written (Moore’s Law). One way in which information is precisely written in the brain is that the molecular signal for the conjunction of presynaptic “read” spikes and postsynaptic “write” spikes, a local calcium ion response, is localized to an individual synapse. However, recent data, as well as basic physics, reveals that this localization is not 100 %, because a small fraction of the calcium leaks to nearby synapses, producing spurious synapse strengthening. This leakage is the Achilles heel of wet, neural, computing, and we suspect the neocortex is, above all, a device for avoiding the potentially catastrophic consequences of such errors. Even very rare errors can snowball as Hebbian learning progresses, especially for nonlinear neurons and higher order statistics, leading to an “error catastrophe”. Just as replication errors impose a universal “speed limit” on Darwinian adaptation, Hebbian inaccuracy imposes a limit on synaptic learning.

How does the neocortex raise the learning speed limit? The basic mechanism (as in accurate DNA copying) may be Hebbian *proofreading*[6] (Fig 2).

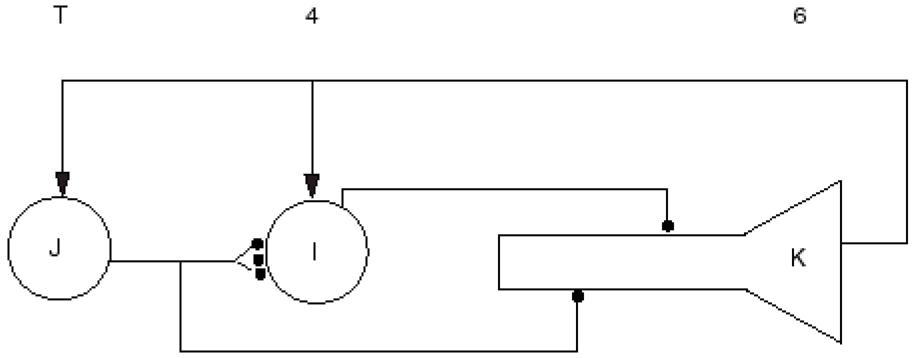

Fig 2. A proposed “canonical neocortical microcircuit” that proofreads synaptic updates and avoids learning error catastrophes. T/J refers to thalamus, 4/I to spiny stellate cells in the thalamorecipient layer of cortex, and 6/K to the coincidence-detecting plasticity-gating deep pyramidal cells.

The Hebbian connection from J to I (representing, for example, a thalamocortical connection) can undergo strengthening as a result of coincident firing of the J and I cells, but this strengthening is not 100% precise, and may lead to a learning catastrophe (weights randomize). This can be prevented using a second independent assessment of coincidence using a special “K” neuron, which then “gates” the plasticity of the feedforward connection. The gating signal is fed to both the input (J) and output (I) cell, and conjunction of pre- and postsynaptic gating signals is required for Hebbian updates to occur. This circuit closely resembles that found in neocortex, but as yet the proofreading hypothesis is unproven.

In all known examples of intelligence (evolution, brains and computers) the key step is writing information accurately. Future intelligent neuromorphic devices will also require accurate “synapses”, and, probably, neocortex-like “proofreading”. It’s likely that rather than using supercomputers to understand the brain[7], we will need to understand the brain to build hypercomputers.

References

1. A. Pouget, Dayan P and R.S. Zemel, *Inference and computation with population codes*, **Annu. Rev. Neurosci**. **26**, pp. 381–410, 2003

2. D. Hubel and T. Wiesel, *Receptive fields,Binocular interaction and functional architecture in the cat's visual cortex.* **J of Physiol. (London)160**, pp. 106-154, 1962

3. R.W. Guillery, *Branching thalamic afferents link action and perception*, **J. Neurophysiol. 90**, pp. 539–548, 2003

4. A. J. Bell and T. J. Sejnowski, The "*independent components" of natural scenes are edge filters*, **Vision Research 37**, pp. 3327--3338, 1997.

5. S.M. Sherman and R.W. Guillery, *Exploring the Thalamus*. Academic Press, San Diego 2001

6. P.R. Adams and K.J.A. Cox (2002). *A new interpretation of thalamocortical circuitry.* **Phil Trans Roy. Soc B. 357** pp.1767-1779, 2002

7. H. Markram. (2006). The blue brain project. Nature Reviews Neuroscience 7, pp.153-160

Paul Adams and Kingsley Cox

Department of Neurobiology, SUNY Stony Brook, Stony Brook, NY 11794, USA
Email: padams@notes.sunysb.edu
http://syndar.org