Intelligence, Sex and the Single Neuron – a Position Paper

Paul Adams and Kingsley Cox SUNY, Stony Brook, NY 11790

Some New Principles – Our unconventional approach to understanding human intelligence is rooted in neuroscience, specifically the assumption that to make progress one must first understand the basic principles underlying neocortical circuitry, which is shared by all mammals. Thus, rather than frontal attack, we advocate siege. Our second unusual premise is that the only well-understood case of biological intelligence, Darwinian evolution, could provide essential clues. By this we do not mean that a combination of genetic algorithms and neural networks could underlie human intelligence, but instead that the underlying logic of these 2 forms of intelligence (indeed, of any form) might be similar. One important clue is that "biology" falls into 2 parts: our understanding of the general machinery common to all life (genes/DNA, proteins, cells etc) and the myriad, concrete, apparently intelligent, solutions to specific problems that different species evince. One could never hope to understand the former by studying the details of the latter, though this is the strategy used by almost all those interested in neural intelligence. But the most important clue is that superaccurate replication is essential for evolution and "Darwinian intelligent design". Thirdly, we assume, conventionally, that learning is key to intelligence. But since the brain is composed of neurons, we think, rather surprisingly, that "intelligence" can emerge even at a single neuron level, in the following sense. The key brain process underlying intelligence must be learning relationships between inputs (including inputs that result from an animal's actions), and thus boils down to learning from higher-order correlations ("hocs"). Ultimately hocs can only drive learning in individual neurons, because the neuron is the only place that many individual inputs come together. If a neuron cannot learn from hocs, the whole enterprise is doomed. We think the mistake everyone makes is to assume that single-neuron hoc learning is trivial, and that therefore intelligence emerges from special conditions on ensembles of neurons. This mistake is analogous to supposing that life is what emerges from special combinations of genes, but instead the emergence of one gene (coding an accurate general selfreplicase) did the trick. Gene assemblies create species, not life.

These principles lead directly to our main conclusion: single neuron hoc learning is only possible when the (basically Hebbian) machinery that updates single synapses is highly connection-specific. We can show that elementary physics precludes synapses from being specific enough to learn from hocs involving nontrivial numbers of inputs (eg hundreds), and therefore the key to intelligence was the appearance of neocortical circuitry that allows the attainment of the necessary specificity. This argument is closely related to that underlying the emergence of modern life: sufficiently accurate DNA replication is achieved by complex non-DNA machinery (notably, proofreading polymerases). This new view of intelligence initially seems completely unrelated to most current effort but leads to a surprising new perspective on the emergence of human-level "mind". We summarize the 3 main strands of evidence for this unorthodox viewpoint, and end with a speculation about human intelligence.

Synapses cannot operate independently, though they must - In the standard connectionist viewpoint the strength of an individual synapse must slide, up (LTP) or down (LTD), in response to the arriving local signals (input and output spikes, together with global signals), and it is optimistically assumed this can be done. But it must be done precisely (by an appropriate amount) and accurately (ideally, independently of changes at other synapses). This cannot be achieved biologically, because synapses are close-packed (sometimes separations < 100 nm), and involved chemicals diffuse. The difficulty is compounded by the requirement that synapses be well coupled electrically. Accuracy for "up" adjustment is provided mainly by spines, but even here 1% (or more) of the essential triggering calcium escapes to other synapses; this limit resembles the 1% error rate intrinsic to unvarnished Crick-Watson basepairing. LTD accuracy is less well understood but likely to be even less. Recent data have confirmed that these processes are not synapse-specific, though this is required by connectionist theories.

HOC learning requires independence - In our view the key (but strangely neglected) question for neural intelligence is whether neurons can, in analog fashion (spikes, calcium etc), learn from the hocs present in their hundreds or thousands of inputs even when "crosstalk" is present (changes in connections are interdependent). If a neuron can so respond, then the remaining issues (e.g. coordinating learning in millions of neurons) become one of detail, for which general principles may be rare. We seek principles at the most fundamental level. We test this with toy models: if robust toy learning is impossible, to hope that the necessary complications and compromises of real neurons can solve fundamental difficulties is not theory, but magic. Our toy version of realistic single neuron hoc learning has 2 parts. First, we represent crosstalk by an error matrix **E**, which describes how updates at one connection directly influence those at others (of course, in Hebbian learning all updates reflect activity at multiple inputs and are thus *indirectly* coupled, as required for correlational learning). If crosstalk does not systematically bias learning, **E** has equal entries along the diagonal (slightly less than one) and very small but equal offdiagonal terms. Second, we use the simplest possible hoc-driven model, ICA. Specifically, we study a 1-unit learning rule with a (correctly-signed) nonlinear Hebb term and an explicit normalizing term; **E** operates on either of these terms. With no crosstalk it is known that this rule can always find an unmixing column of an orthogonal mixing matrix, provided the relevant source has a nonGauss distribution. Orthogonality guarantees learning, but is not biologically plausible: real circuits can only approximately pairwise-decorrelate inputs. We find that if the inputs are not sufficiently white, hoc learning fails even in the absence of crosstalk: now the neuron's learning is only driven by second-order correlations (socs). Most significantly, even in those cases where inputs are sufficiently well-whitened that hoc-learning succeeds in the absence of crosstalk, it often fails when crosstalk exceeds a biologically plausible (indeed, inevitable) level, and defaults to approximate soc learning. There are 2 capital points: only (biologically-unobtainable) perfect whitening immunizes against crosstalk, and hoc learning fails *completely*, and defaults to (approximate) soc learning above a crosstalk threshold. Since socs cannot provide insight into the causes of experience (the hallmark of "understanding"), we conclude that the necessary, and hitherto overlooked, ingredient for neural intelligence is attainment of adequate synapse independence. Similar "base-independence" is the key to life, the other known form of intelligence. We do not claim the brain does ICA, merely that neurons must be able to learn from hocs; the (crosstalkless) ICA model is the only case this is guaranteed.

Neocortical circuitry can perform the necessary Hebbian proofreading operation - Biophysics precludes achievement of the synaptic independence required for the type of massively parallel, analog, learning underlying intelligence, yet at least some mammals are intelligent. We therefore suggest that "canonical" circuitry which underlies all neocortex in all mammals allows great improvement in effective synapse independence. This circuitry is, by definition, distinctive and essential to neocortex, and thus we focus on thalamocortical interactions, which essentially define neocortex (other features, touted as "canonical", are not unique to neocortex). The key idea here

is "proofreading" (analogous to the machinery underlying high DNA replication accuracy): crosstalk implies that spikepairs at other synapses can perturb adjustments of a given synapse in response to local spike-pairs (pre-post or post-pre), resulting in "errors". Proofreading amounts to making a second, independent, measure of local spike pairing; since the errors in the 2 mirror pair-detection operations are independent, if both operations must agree to generate synapse-adjustment, the 2 error rates multiply together, vastly increasing accuracy.
The requirements for the actual implementation of this scheme are rather stringent, but they happen to correspond to many hitherto mysterious features of thalamocortical circuitry and physiology. Specifically, the scheme requires that a set of neurons (in layer 6) receive inputs from both thalamocortical afferents and their midlayer (eg 4) targets), act as post-pre coincidence detectors, and gate (probably presynaptic) LTD of the connections they "monitor" (postsynaptic LTP accuracy would derive from spines). This must be achieved by sending rapid modulatory signals to both sides of the relevant connections (i.e. to a relay and its midlayer cortical target). While this apparatus seems extraordinarily complex and tricky, its implausibility largely evaporates when one realizes that it largely corresponds to the actual, known, machinery, which has hitherto eluded coherent explanation. Our hypothesis is extraordinary, but it matches extraordinary facts. There is one interesting and apparently fatal objection to "Hebbian proofreading": it seems to postulate a monitoring "proofreading" neuron for every (feedforward, thalamocortical) connections. However, provided that the expected time between spike coincidences is long compared to the operation of the proofreading machinery (~ 100 msec), a "distributed" scheme should work well: a layer 6 corticothalamic neuron can successfully monitor all the connections made onto a given cortical target. Even if this condition is not always met, such proofreading errors merely lower overall accuracy towards the floor provided by regular Hebbian learning, and, crucially, always lead to significant gains in the numbers of inputs from which hocs can be learned.
Sex in evolution and symbolic intelligence - We believe the key to neural intelligence is that single neurons can successfully detect and learn from the hocs present in their hundreds or thousands of inputs, essentially by an analog operation, which defeats the dimensional curse and gives a compact description of inputs statistics as the stable fixed point of a nonlinear Hebbian learning rule. We suspect that this lurks at the heart of all accounts of neural or neurally-inspired intelligence. But all other authors fudge the difficulties attendant on this key operation, and instead focus on clever strategies for harnessing the posited operation, in the service of general, or more usually specific, tasks. This is like mistaking the power of the Turing-von Neumann computer for that of the programs that run on it. We argue that the core analog operations (e.g. massively parallel multiplication of many inputs values) are essentially miraculous, because they cannot be done by real biophysical machinery. One possible strategy to deal with this is to impose suitable approximations to describe the relevant input statistics, but we think this is the wrong approach: it would be far more powerful to reduce biophysically inevitable approximations, primarily synapse coupling, This is essentially the strategy that Darwinian evolution employs: it throws immense resources at reducing the replication error rate (to the current quasimiraculous level of ~ 10^{-10}), and leaves the detail to the running of the replication/selection process (which in turn maximizes parallelicity in endless particular ways). One does not really need to worry too much about how exactly organisms photosynthesize or excrete. Though these are interesting

questions, there is no hope that one will throw light on the other. Biologists recognize that both have emerged as a result of repeated rounds of highly accurate replication.
Thus although our analysis is rooted in technical analyses of connectionist networks, synaptic biophysics and neocortical circuitry, it leads to a breathtakingly broad picture of intelligence: it would be the result of selective (i.e. accurate) amplification of essentially unlimited distinct states. The difference from conventional ideas about "selforganization" is the paramount emphasis on the physical mechanisms underlying selectivity. We claim one must have the right physical apparatus to be intelligent: at least one neuron equipped with proofreading. An obvious problem remains: humans also have the same canonical neocortical machinery, yet seem to be much smarter than other mammals. Can our new viewpoint throw light on this difficulty? Here our ideas become more speculative. We start again from the key analogy, between biological and neural, intelligence. Clearly, replication/selection, while the main dynamo of evolution, is rather limited. Essentially, the mutation rate (which is capped by required replication accuracy, following Eigen) limits the rate at which a population can evolve in response to environmental change (we remark that "environment" is just hocs-in-disguise: it constitutes a set of inputs to the population, which defines the replication rates of all possible sequences; without epistasis there is no hoc-sensitivity). Evolution gets stuck at the prokaryotic level. The solution is sex, in essence a shared protocol for the exchange of genetic information. Mutation leads to the appearance of near-neutral alleles, which are recombined by sex. Note that sex without mutation is possible, but soon runs through allele combinations. Mutation essentially never runs out of novelty, but cannot occur fast enough to cope with a shifting world. One gets the best performance using both (as GA practitioners know). But stating this does not constitute a theory of intelligence: it merely sets the scene for the real work, understanding the logic and machinery. Sex is symbolic: the set of alleles constitutes a vocabulary that implicitly describes the world: these are sequences that, individually, survived the selection process. The point of sex is that combinations of these symbols may survive better than one would predict from the survivability of individual symbols (much as semantically meaningful sentences are a small fraction of possible sentences: they correspond to events that could happen according to one's current understanding). Language moves powerful but frail unsupervised hoc learning towards trivial, supervised, crosstalk-resistant, learning (because population-level evaluation operates on semantically independent symbols). In this preliminary picture, human intelligence would have 2 intertwined components, both rooted in specific physical machinery. First, in common with other mammals, isolated individual humans would be able to "understand" simple aspects of the world, by discovering hoc patterns using superaccurate synapses. This yields "insights" but they are fallible because there is no way for the individual to know that hoc learning has succeeded rather than defaulting to soc learning (they are both stable fixed points of a learning rule). Such insights emerge from a massively parallel, highly accurate, analog computation in the neocortex. Candidate insights assemble, as a result of symbolic communication ("neural sex"), into combinations that successfully match the world (stable fixed-points of population dynamics that combine language and real-world testing). It is the yoking of the symbolic (sexual, allelic) and insightful (asexual, mutational) processes that yields human levels of intelligence. Of course these ideas are fuzzy, and reminiscent of conventional thinking. This account links one unsatisfactory debate (eg symbolic/connectionist) to another (role of sex). But this is also a strength, since progress in the latter, rather concrete, debate would yield progress in the former.