Hocus-Socus: An Error Catastrophe for Complex Hebbian Learning Implies Neocortical Proofreading

Cox, Kingsley J.A.[1,2] and Adams, Paul R.[1,2]

[1]Dept Neurobiology, Stony Brook University, NY
[2]Kalypso Institute, Stony Brook, NY

Correspondence to: P.R. Adams[1] Correspondence and requests for materials should be addressed to P.R.A. (email: padams@notes.sunysb.edu)

The neocortex is widely believed to be the seat of intelligence and "mind". However, it's unclear what "mind" is, or how the special features of neocortex enable it, though likely "connectionist" principles are involved *[A]. The key to intelligence[1] is learning relationships between large numbers of signals (such as pixel values), rather than memorizing explicit patterns. Causes (such as objects) can then be inferred from a learned internal model. These relationships fall into 2 classes: simple pairwise or second-order correlations (socs), and complex, and vastly more numerous, higher-order correlations (hocs[B]), such as the product of 3 or more pixels averaged over a set of images. Thus if 3 pixels correlate, they may give an "edge". Neurons with "Hebbian" synapses (changing strength in response to input-output spike-coincidences) are sensitive to such correlations, and it's likely that learned internal models use such neurons. Because output firing depends on input firing via the relevant connection strengths, Hebbian learning provides, in a feedback manner, sensitivity to input correlations. Hocs are vital, since they express "interesting" structure[2] (e.g. edges), but their detection requires nonlinear rules operating at synapses of individual neurons. Here we report that in single model neurons learning from hocs fails, and defaults to socs, if nonlinear Hebbian rules are not sufficiently connection-specific. Such failure would inevitably occur if a neuron's input synapses were too crowded, and would undermine biological connectionism. Since the cortex must be hoc-sensitive to achieve the type of learning enabling mind, we propose it uses known, detailed but poorly understood circuitry and physiology to "proofread" Hebbian connections. Analogous DNA proofreading allows evolution of complex genomes (i.e. "life").

This view, combining insights from synapse biophysics, molecular evolution, neocortical anatomy and neural learning theory, seems as unpromising as the notion that life is the outcome of amplified molecular accidents, to which it is closely linked[C]. Recent data suggest that Hebbian adjustments are highly[3], but not completely[4,5] specific, because of excellent (~99%) confinement of calcium[6,7] and its effects[8] by spines[D]. Since biological processes are usually error-tolerant the observed specificity might suffice for learning hocs, but this has never been tested, and there is a highly relevant case where extraordinary accuracy is essential, DNA replication. Darwinian evolution, a type of chemical complex learning from the world[9], is only possible because error rates for base

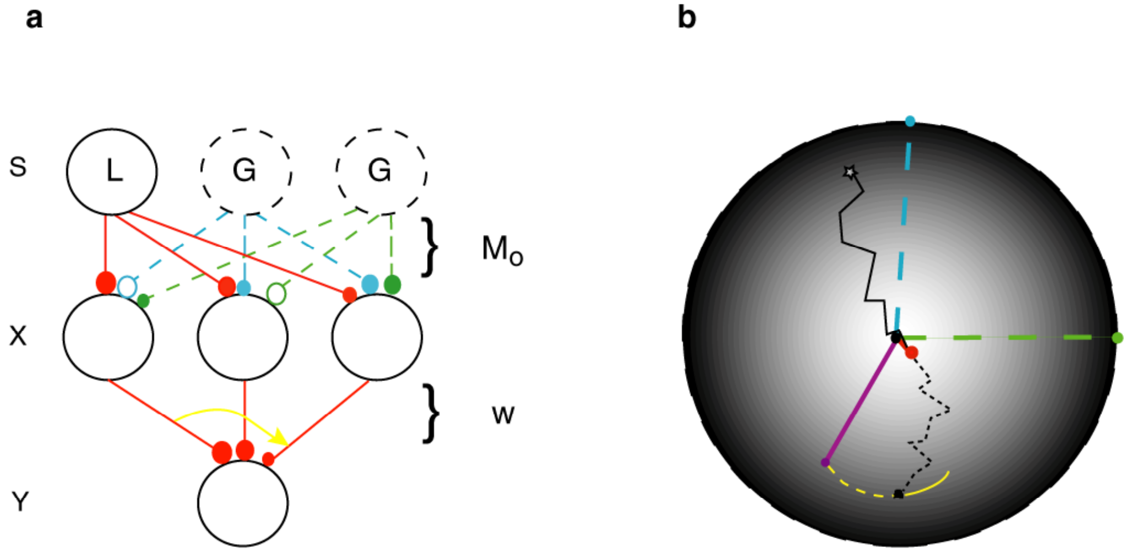
*superscripted letters refer to Supplementary Notes; the Supplement also contains additional material.

copying are comparable to reciprocal genome lengths[E,9,10]. Accurate replication (per-base error rates $<10^{-9}$) is achieved by multiple mechanisms[11] which evolved in a series of information-enriching transitions[12]: relatively (~99%) selective base-pairing; selectivity conferred by replicases; proofreading; mismatch repair. The largest contributor is proofreading, reducing the error rate from $< 10^{-3}$ to $<< 10^{-6}$, since 2 independent pairing events must concur.

We proposed[F,13,14,15] that key neocortical circuitry accomplishes a conceptually-identical proofreading operation on 2 independent measures of spike-pairing, allowing large improvements in Hebbian accuracy, and otherwise usually impossible feats of learning. We now show that the required circuitry closely matches recent data[16] and describe computational results providing the crucial missing link: complex learning by a model neuron typically collapses to simple learning, if Hebbian specificity falls below a threshold comparable to that expected (and observed) from biophysics. This test requires a model where learning (i.e. convergence to stable weights) depends on input correlations generated in a defined manner, and on a nonlinear Hebb rule. Independent Component Analysis (ICA)[17,18] meets these requirements. This model learns implicit target weights using the hocs in an ensemble of input patterns. We used the simplest possible "crosstalk" model[G,19], corresponding to the usual genetic assumption of base- and position-independent copying-accuracy, though we obtained similar results when "hotspots"[F] were introduced. As in Darwinian evolution[9,10,E], the threshold depends on the learning task[H], but typically falls within a biophysically-plausible range, so if the neocortex is to solve diverse problems, it must have wetware overcoming this limitation.

Complex learning power reflects the number of inputs whose hocs can be exploited by Hebbian rules, and is therefore best done in individual neurons, rather than dendritic segments[I]. Our model is based on single-unit ICA[J,17,20], a minimal hoc-based abstraction of object identification. Input vectors $\mathbf{x}$ (with pixel-like elements x,) generated by linearly combining n independently-fluctuating unknown object–like "sources" using an approximately orthogonal[K] square "mixing" matrix $\mathbf{M_0}$ are applied to the adjustable weights w of a neuron whose output y (the inferred "object") is the weighted input sum $\mathbf{x^T w}$ (Fig.1a). x and y are mean rates rather than detailed descriptions of firing times which may be necessary to predict real neuron output, since this "connectionist" model doesn't respond to temporal sequencing. Timing would make it even more difficult for real synapses to achieve high specificity[L]. The ith weight adjustment is made using the nonlinear Hebbian rule $\Delta w_i = +/- \text{ k } x_i \text{ f}(y)$. k is a small learning constant, and f any sufficiently smooth nonlinearity; we usually used the statistically robust[17] tanh which for typical superGaussian sources requires a negative sign ("antiHebb") in the rule[17,20]. In real neurons this multiplication could be implemented by spike coincidence detection[M]. Linear Hebb rules are only sensitive to pairwise correlations[19,21]; nonlinearity provides additional sensitivity to hocs[O]. Hebbian rules produce weights that grow or shrink without limit, and require stabilization: we divided the weight vector by its new length after each adjustment[20]. Similar "normalization" could be achieved by a variety of mechanisms and is "multiplicative", confining the weight vector to a unit sphere[22] (Fig. 1b).

The 1-unit rule also requires that inputs be preprocessed, or "whitened", to remove socs; we found that partial whitening typically sufficed[P]. A random **M** was used to generate an initial batch (typically 10^3) of mix vectors, for which a small-sample covariance matrix $\mathbf{C_s}$ was calculated[P]. $\mathbf{M_o}$ was formed using $\mathbf{M_o} = \mathbf{C_s}^{-1/2}\mathbf{M}$, so $\mathbf{C_L}$, the large-sample (10^5) covariance matrix of the imperfectly decorrelated, "off-white", mix vectors, is close to a scaled identity matrix **I** (Fig. 1c), to an extent that depends on the small-sample size. In practice perfect decorrelation cannot be achieved using reasonable samples, or with biological crosstalk and finite k[Q19]. If the vector **w** converges to a row of $\mathbf{M_o}^{-1}$, the output tracks a source; to simplify model learning and its interpretation, in most cases all sources but one were Gaussian so only one equilibrium, extracting the nonGaussian source, is stable[R,17,20,23] (Figs. 1b, 2a,b).



c

$$\mathrm{M_O} = \begin{bmatrix} 0.67 & -0.65 & 0.26 \\ 0.61 & 0.35 & -0.70 \\ 0.36 & 0.68 & 0.64 \end{bmatrix} \quad \mathrm{M_O}^{-1} = \begin{bmatrix} 0.72 & 0.61 & 0.37 \\ -0.66 & 0.34 & 0.65 \\ 0.30 & -0.71 & 0.65 \end{bmatrix} \quad \mathrm{C_L} = \begin{bmatrix} 1.88 & -0.01 & -0.06 \\ -0.01 & 1.97 & 0.02 \\ -0.06 & 0.02 & 1.99 \end{bmatrix} \quad \mathrm{E} = \begin{bmatrix} 0.88 & 0.06 & 0.06 \\ 0.06 & 0.88 & 0.06 \\ 0.06 & 0.06 & 0.88 \end{bmatrix}$$

Figure 1 The ICA-with-crosstalk model: structure, behaviour, parameters

a shows a model neuron (output y) receiving input from 3 mix signals x via adjustable connection weights w. The mix signals are formed by combining 3 independently and symmetrically fluctuating sources s via a set of fixed mixing coefficients (different size colored dots; open dots are negative), the elements of the almost orthogonal matrix $\mathbf{M_o}$. In practice ICA is done in 2 stages: initial linear PCA ("whitening") followed by nonlinear learning; these are combined in this figure by replacing **M** by $\mathbf{M_o}$. The first column of $\mathbf{M_o}$

corresponds to the red coefficients, arising from the nonGaussian (typically Laplacian) source shown as a solid circle ("L"). The other sources are Gaussian (dotted circles; "G").

b diagrams schematically the 3 weights, zigzagging (solid: subthreshold crosstalk; dotted: suprathreshold) under the influence of successive patterns, and confined by normalisation to the unit sphere. For low crosstalk weights zigzag to the IC (red dot), above threshold to the approximate PC (solid yellow line). The 3D weight surface has been rotated so the direction of the red column of $\mathbf{M_o}$ points straight at the reader, so the direction of the first row of $\mathbf{M_o}^{-1}$ (to which it is almost parallel) points almost to the reader (short solid red line starting at the black dot origin and terminating on the sphere as a red dot). The other directions of almost orthogonal rows of $\mathbf{M_o}^{-1}$ are also shown as blue and green dotted lines. The red dot is the target weight vector that allows the neuron to track the nonGaussian source (the "IC"). The purple line shows the least PC, and the yellow line the loci of the terminations of the least eigenvectors of **EC** on the sphere (i.e. the stable weights obtained using purely Gaussian sources at various errors). Just suprathreshold error triggers a movement (dotted zigzags) from the approximate IC to the square; further increase in error moves the learned average weights along the solid yellow line; the dotted yellow line is not stable when sources are nonGaussian.
c shows the mixing matrix $\mathbf{M_o}$ and its inverse, the unmixing matrix $\mathbf{M_o}^{-1}$; the red row is the only stable IC and corresponds roughly to the red coefficients in the first column of $\mathbf{M_o}$. Since $\mathbf{M_o}$ is only approximately orthogonal, the covariance matrix $\mathbf{C}_L$ of even a large (100,000) batch of **x** has offdiagonal elements very small but nonzero, and slightly unequal diagonal elements. The error matrix **E** (which has equal diagonal elements Q and equal offdiagonal elements (1-Q)/(n-1)) is shown with entries corresponding to the threshold in Fig. 2. For further detail see Supplementary Legends.

We introduced crosstalk by modifying the rule to $\Delta\mathbf{w} = +/-k\mathbf{E}\ \mathbf{x}\ f(y)$, where **E** is a symmetric "error matrix" assigning a fraction (1-Q) of an adjustment to the other weights, dividing it up according to the offdiagonal elements of **E**[S,T,19]. Zero crosstalk, assumed in standard models, implies Q ("quality") = 1. Usually we set offdiagonal elements of **E**, and also diagonal elements, to be equal (Fig. 1c), corresponding to the standard connectionist assumption that all connections of a given type are equivalent, and to spatiotemporal averaging of varying synaptic configurations[T,19].

In most tests the nonGaussian source had a Laplacian distribution[U]. With zero error the rule converged to the weights corresponding to this source, the "IC" (Figs. 1, 2). A low level of error ("crosstalk", expressed as a per-connection quantity that is independent of n[T,19]) produces only slight degradation of learning, but, crucially, above a narrow threshold range, weights snap from the IC to a new average direction[V]. If the nonlinear rule fails to learn from hocs above a threshold, this new direction could correspond to mere soc learning. We tested this using the same $\mathbf{M_o}$ but with all sources Gaussian, so the mix vectors exhibit only socs[W]. Now the error-free nonlinear rule learned the least eigenvector of $\mathbf{C}_L$, as expected for an antiHebbian rule[P]. As crosstalk increased, the learned vector gradually moved away from this direction, and above the nonGaussian threshold the weights learned for either mixed Laplacian-Gaussian sources or pure Gaussian sources were identical (Figs. 1b, 2a,b). Furthermore, the learned vector for Gaussian sources tracked the expected theoretical curve[19] (corresponding to the least eigenvector of **EC**) for a linear rule (Fig. 2a), although the rule is nonlinear. Crucially, minor crosstalk makes the nonlinear rule behave linearly, ignoring hocs. This was true for different **M**s (though the threshold varied; for 4 cases studied in detail the average threshold was 0.04 +/- 0.03 (SD)), and for different source distributions or degrees of whitening (being more error sensitive for lower kurtosis sources or less whitening[X]).

We are not proposing the brain does ICA, though it may do something similar[Y,24,25,26]. Instead our results suggest a principle: nonlinear Hebbian rules become insensitive to hocs above a threshold crosstalk level[Z]. A normalized nonlinear correctly-signed rule automatically learns ICs if inputs are generated by square linear mixing. If inputs are generated differently, for example by rectangular mixing, nondeterministically or nonlinearly, a single neuron may not learn any stable weight vector[27], but if it does, enough crosstalk will cause failure[AA].

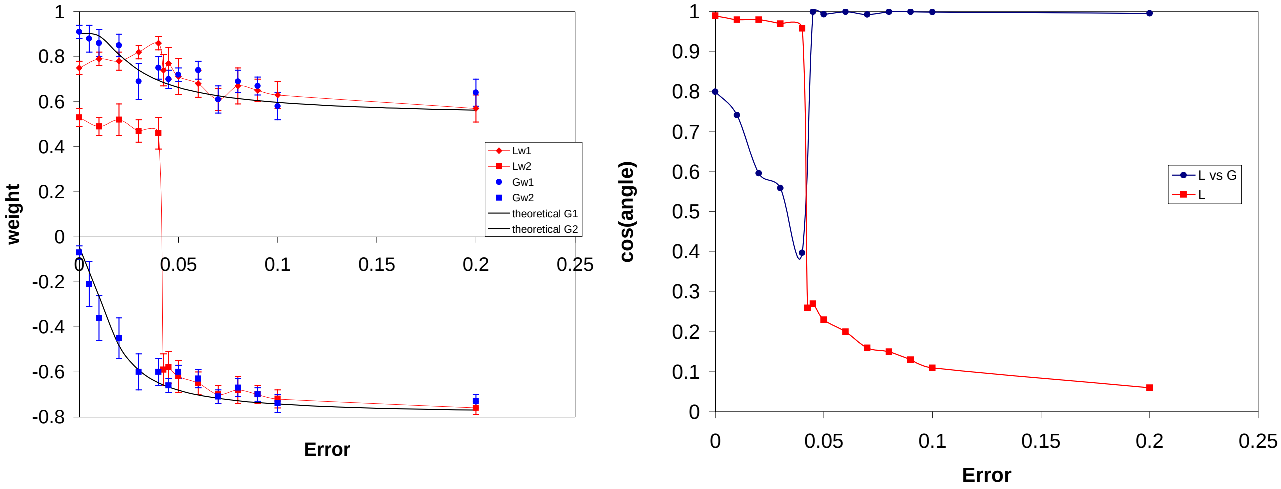


Figure 2 Crosstalk causes hoc learning to collapse to soc learning in 1-unit ICA.

Calculations were done using the conditions in Fig. 1, with 3 inputs and weights, using f(y)= tanh(y), explicit normalisation and antiHebb learning. The left hand plot shows two of the steady-state averaged weights using 1 Laplacian source (red) or all Gaussian sources (blue; error bars show SD). At the threshold per connection error b = 0.0425 the Laplacian weights snap to match the all-Gaussian weights. The per connection error b is related to Q (Fig. 1) by b = (1-Q)/nQ, and expresses the expected dependence of crosstalk on biophysical parameters[19]. The black lines show theoretical weights calculated from the least eigenvectors of $\mathbf{EC}_L$[19]; $\mathbf{C}_L$ was estimated from a sample of 100,000 input vectors. The right hand plots show the cosine of the angle between the Laplacian and Gaussian weight vectors as a function of error (blue line); at the critical error the Laplacian weight vector jumps to the Gaussian weight vector. The red line show how the cosine of the angle between the weight vector and the first row of $\mathbf{M_o}^{-1}$ (Laplacian source) changes with error (again with sharp change at b = 0.0425). See Supplementary Legends for details. k = 0.002; similar results were obtained with k = 0.0002

Why does crosstalk prevent learning from hocs[BB]? A weight vector parallel to a row of $\mathbf{M_o}$ is a stable equilibrium of an accurate averaged nonlinear rule[17,20] but the rule may have other equilibria. For a linear rule and any nonwhite input distribution the eigenvectors of **C** (Principal Components; PCs) are equilibria, and the greatest (or for antiHebb, least) is, typically, stable[19,21,P]; this should also be true for nonlinear rules with Gaussian inputs. We confirmed that the leading or (for antiHebb rules, least) eigenvector of **C** is stable for Gaussian inputs, even for a cubic nonlinearity with no linear term. However for nonlinear rules, sufficiently nonGaussian inputs destabilize this PC, and stabilize the IC[17,20]. Suprathreshold error apparently nulls the nonlinearity, destabilizing the IC and restabilizing the approximate PC (least or largest eigenvector of **EC**), because error moves the equilibrium weights slightly away from the IC, eventually invalidating the stability proof[17,20]. Further study should reveal what factors other than source kurtosis and orthogonality[X] of $\mathbf{M_0}$ (e.g. $\mathbf{M}$, n and bit resolution) set the threshold and its sharpness, and why. However, either these factors reflect the task specifics (and cannot be circumvented) or unavoidable neural limitations.

These results suggest that individual neurons typically cannot learn connection weights that reflect input hocs unless the necessarily nonlinear Hebbian rule is highly (~95%) accurate[T]. Indeed, they underestimate the problem since they ignore additional effects of crosstalk on biological prewhitening[T,19]. Even if only ~1% of the calcium entering a spine escapes to the shaft[7] but there are 10 or more synapses within range of that calcium[28,29], such accuracy may be unobtainable[19CC]. Crowded synapses are inevitable if neurons learn from many inputs[W]. Since neocortical neurons manifestly do learn from hocs, even though their numerous inputs may obey more less favourable statistics than for ICA, the neocortex must presumably use a non-synaptic (neuronal) strategy for increasing Hebbian accuracy[13,14,15]. The root problem is coincidence-detection failure: because of intracellular messenger diffusion from nearby synapses experiencing coincidences, a connection may register a "false spike-pair", analogous to incorrect base-pairing. The obvious way to overcome this uses an additional, independent measure of near-coincident firing of input and output neurons contributing to the synapse; double mistakes should be rare. We have suggested[13,14,15] the neocortex might contain (in layer 6) dedicated "Hebbian neurons" detecting coincidences across connections, using branches of the relevant axons, and supplying these independently detected coincidences in near real-time (<100 msec[DD]), to the relevant (probably thalamocortical) connections, so if the "second" (neuronal) coincidence confirms a "first" (synaptic) coincidence, the relevant weight is allowed to change. Fig. 3a diagrams the necessary wetware. Selective confirmation delivery to the relevant connection could be achieved by applying it pre- and postsynaptically (e.g. to a relay and its layer 4 target), requiring that both sides of the connection receive it (Fig. 3a). This "proofreading" strategy would seem to need a dedicated proofreading neuron for every anatomical feedforward connection (recurrent connections may not need proofreading if they learn socs) even if comprised only of silent synapses. However, since coincidences across connections are probably rare (antiHebbian learning and NMDAR maturation tend to reduce them), a proofreading neuron could monitor many connections in a distributed manner (Fig. 3b): while it could wrongly enable strength changes at connections not experiencing genuine coincidences, this would vanish in the sparse coincidence limit (much as interference in associative memories vanishes for sparse patterns). The diagram in Fig. 3b (see also Supplement Fig. 1), which goes beyond our previous sketch[15], matches known but mysterious "universal" thalamocorticothalamic circuitry and physiology[30] , and could form the backbone of the cortical "column". There is remarkably close agreement between these requirements for distributed proofreading and recent counterintuitive data on the pattern of CT feedback[16,EE].

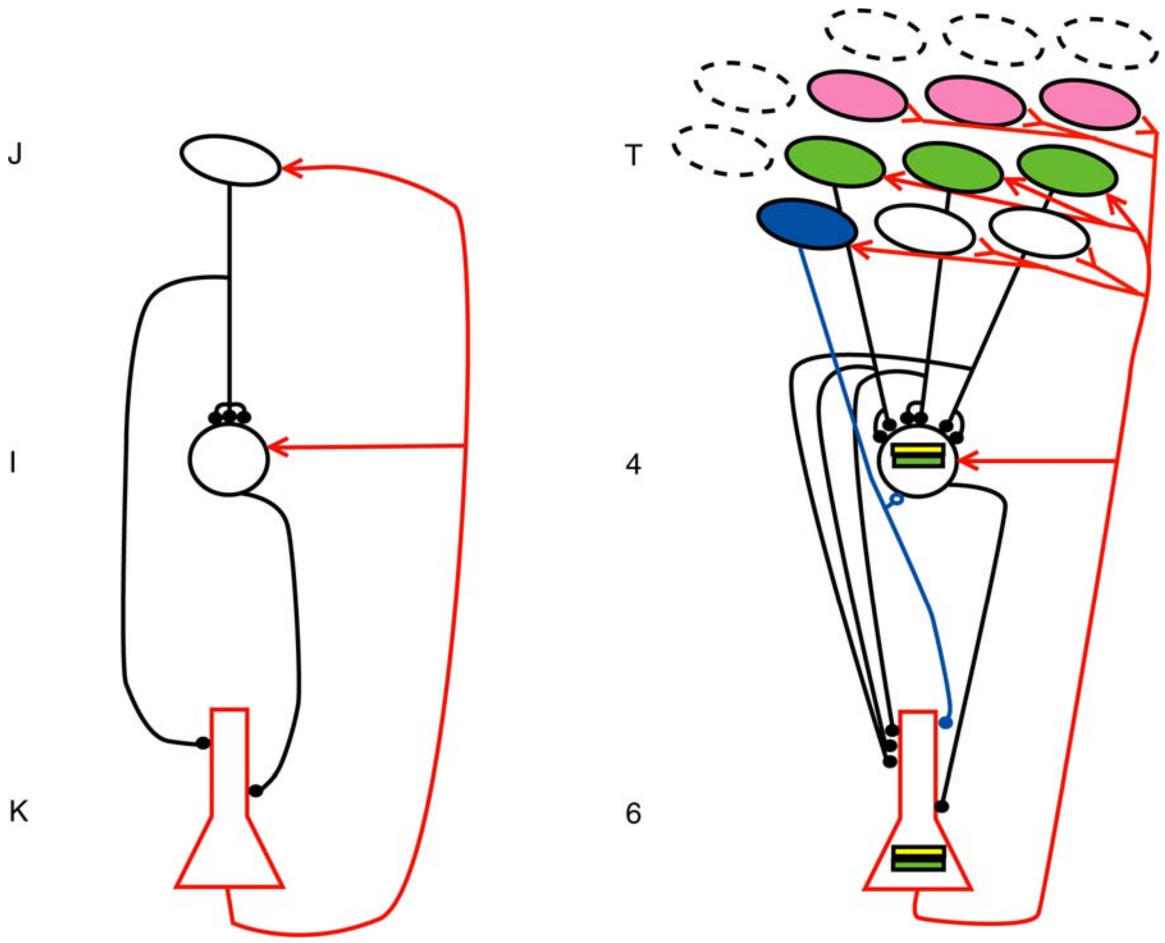


Figure 3 Proposed thalamocortical circuitry for dedicated or distributed proofreading.

Fig. **3a** (Left) Dedicated Proofreading. An input J connects to a neuron I and also, via a weaker connection, to a dedicated proofreading partner K. The K cell also receives weak input from the I cell. K feeds back to both J and I via modulatory connections (red arrows). K fires when the J and I cells fire near-coincident spikes and shifts the target J-I connection to "plasticity approved" mode by conjoint modulation of the input and output sides.

3b. (Right) Distributed "PushPull" Proofreading. The diagram makes the identification J = thalamic relay (T), I = layer 4, K = layer 6 CT cell and shows one possible version of distributed proofreading, for concreteness drawn for an orientation-tuned layer 4 simple cell (i.e. responding to a horizontal edge) ; only off relays, which generate the green off-lobe of the layer 4 RF, are shown; the green, "overlapping' and "matching"[16], relays contribute to the off-lobe; the on-lobe is shown yellow, and the corresponding "overlapping" but "nonmatching" [ZZ] off-relays are shown pink – these do not connect to the layer 4 or 6 cells shown). The layer 6 cell firing modulates its partner layer 4 cell directly to briefly enable thalamocortical plasticity postynaptically. It also modulates the set of thalamic relays (green and blue ovals) that innervate (by silent or nonsilent synapses) its partner layer 4 cell, via a TRN inhibitory cell (omitted), which shifts relays to burst mode, briefly enabling thalamocortical plasticity presynaptically (red arrows). Both pre- and post-enabling are required for the strength change triggered by T-4 spike-coincidence to be expressed; such dual-enabling occurs if the 6–cell rapidly confirms the spike-coincidence "seen" by the relevant thalamocortical synapses. Enablement should be executed before the typical arrival of the next coincidence (~100 msec -10 sec). The dotted ovals correspond to "unavailable" relays that cannot reach the dendrites of the illustrated layer 4 cell. This is a "functional" diagram; see Supplement for an anatomical diagram showing the intervening TRN cell, which innervates all the nondotted relays.

Currently unconnected "incipient" relays[14] , including "nonmatching" relays (pink) and nonoverlapping, open undotted relays , that could form synapses on the 4-cell receive direct depolarizing modulation (reversed red arrows) which maintains them in tonic, plasticity-disabled, mode (unless they receive enabling signals from other 6-cells monitoring the connections they do form, on other 4-cells). Some "nonoverlapping" connected relays (e.g. blue oval), make only silent synapses (open blue dot) and therefore do not contribute to the receptive field. These silent connections must be monitored and should receive enabling input. The scheme closely fits recent results[16,EE]. See Supplement for details.

Our results also suggest a generalization of Eigen's "error threshold"[9,10] (setting the maximum size of genomes) to other forms of learning: learned information depends on the reciprocal of the learning error rate. This seems true for socs[FF,19]. For hocs, the learned information at zero error, the product of the vector dimension (n) and the *w* bit resolution, evaporates at the threshold. Thus the effect of error on soc and hoc learning is quantitatively the same but qualitatively different, being gradual (and tolerable) for the former and abrupt (and catastrophic) for the latter[FF].

This explanation of the neocortical basis of sophisticated learning by individual neurons, the key to intelligence and "mind", is simple, and parallels that accepted as the key to "life"[GG]. In intelligent brains neurons must learn from hocs; perfect Hebbian synapses could accomplish this, but in practice crosstalk usually makes this impossible. A "proofreading" mechanism, conceptually identical to that allowing the evolution of complex genomes ("life"), would allow such learning and matches known, but enigmatic, thalamocortical anatomy and physiology. Nevertheless, even with proofreading, cortical neurons could probably only handle around 1000 inputs (as typically observed), since otherwise synapses become so crowded that crosstalk would increase to the point where hoc learning fails[CC]. This would vastly restrict the learning power of neurons and brains. Evolution may provide useful analogies for understanding learning and intelligence. Perhaps further major evolutionary transitions after DNA/protein[12] provide useful clues about mechanisms enabling human levels of mind[HH,II].

References

1 Poggio, T. & Bizzi, E. Generalization in vision and motor control. *Nature* **431**, 768-774 (2004)

2 Field, D. J. What is the Goal of Sensory Coding? *Neural Computation* **6**, 559-601 (1994)

3 Matsuzaki, M., Honkura, N., Ellis-Davies, G.C. & Kasai, H. Structural basis of long-term potentiation in single dendritic spines. *Nature* **429,** 761-766 (2004)

4 Engert, F. & Bonhoeffer, T. Synapse specificity of long-term potentiation breaks down at short distances. *Nature* **388,** 279-284 (1997)

5 Harvey, C.D. & Svoboda, K. Locally dynamic synaptic learning rules in pyramidal neuron dendrites. *Nature* **450,** 1195–1200 (2007)

6 Yuste, R. Majewska, A. & Holthoff, K. From form to function: calcium compartmentalization in dendritic spines. *Nat. Neurosci.* **3**, 653-659 (2000)

7 Sabatini, B.S., Oertner, T. & Svoboda, K. The life-cycle of Ca^{2+} ions in dendritic spines. *Neuron* **33**, 439-452 (2002)

8 Lee, S.-J.R., Escobedo-Lozoya, Y., Szatmari, E.M. & Yasuda, R. Activation of CaMKII in single dendritic spines during long-term potentiation. *Nature* **458**, 299-304 (2009)

9 Adami, C. *Introduction to Artificial Life.* (Springer-Verlag, 1998)

10 Eigen, M., McCaskill, J. & Schuster, P. The molecular quasispecies. *Adv. Chem. Phys.* **75**, 149-163 (1989).

11 Kornberg, A. & Baker, T.A. *DNA Replication*. (University Science Books, 2005)

12 Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution.* (New York: Oxford University Press, 1997)

13 Cox, K.J.A & Adams, P.R. Implications of synaptic digitisation and error for neocortical function. *Neurocomputing* **32**, 673-678 (2000).

14 Adams, P. & Cox, K.J.A. A new view of thalamocortical function. *Phil. Trans. R. Soc. Lond. B.* **357**, 1767-1779 (2002).

15 Adams, P.R., & Cox, K.J.A. A neurobiological perspective on building intelligent devices. *The Neuromorphic Engineer* 3: 2-8 (2006). Available: http://www.ine-news.org/view.php?source=0036-2006-05-01

16 Wang, W., Jones, H.E., Andolina, I.M., Salt, T.E. & Sillito, A. Functional alignment of feedback effects from visual cortex to thalamus. *Nat. Neurosci.* **9**, 1330-1336 (2006).

17 Hyvärinen, A., Karhunen , J. & Oja E. *Independent Component Analysis*. (Wiley Interscience, 2001).

18 Bell, A.J. & Sejnowski, T.J. An information maximization approach to blind separation and blind deconvolution. *Neural Computation* **7**, 1129–1159 (1995)

19 Radulescu, A.R., Cox, K.J.A. & Adams, P.R. Hebbian Errors in Learning: An Analysis Using the Oja Model. *J. Theor. Biol.* **258,** 489-501 (2009)

20 Hyvarinen, A. & Oja, E. Independent component analysis by general non-linear Hebbian-like learning rules. *Signal Processing* **64,** 301–313 (1998)

21 Oja, E. A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15,** 267-273 (1982)

22 Miller, K.D. & MacKay, D.J.C. The Role of Constraints in Hebbian Learning. *Neural Computation* **6,** 100-126 (1994)

23 Rattray, M. Stochastic trapping in a solvable model of on-line independent component analysis. *Neural Computation* **14,** 421-435 (2002)

24 Bell A., & Sejnowski, T. The 'independent components' of natural scenes are edge filters. *Vision Research* **37,** 3327–3338 (1997)

25 Olshausen, B.A. & Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* **381,** 607-609 (1996)

26 Cooper, L.N., Intrator, N., Blaise, B.S. & Shouval, H.Z. *Theory of Cortical Plasticity.* (World Scientific, 2004)

27 Dayan, P. & Abbott, L.E. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. (Cambridge MA; MIT Press, 2001)

28 Noguchi J., Matsuzaki M., Ellis-Davies, G.C.R. & Kasai, H. Spine-Neck Geometry Determines NMDA Receptor Ca^{2+} Signaling in Dendrites. *Neuron* **46,** 609-622 (2005)

29 Harris, K.M. & Stevens, J.K. Dendritic spines of rat cerebellar Purkinje cells: Serial electron microscopy with reference to their biophysical characteristics. *J. Neurosci*. **8,** 4455-4469 (1988)

30 Sherman, S.M. & Guillery, R.W. *Exploring the Thalamus*. (Academic Press, 2001)

Acknowledgements

We thank the following for comments on a draft of this paper: H. Barlow, P. Dayan, J. Diamond, S.Y Ge, C. Gilbert, G. Hinton, C. Jahr, D. Johnston, A. Maffei, R. Malenka, M. Maravall, H. Markram, D. McKinnon, S. Nelson, R. Nicoll, E. Oja, D. Purves, R. O'Reilly, M. Rattray, M. Sherman, S. Siegelbaum, G. Stuart, C. Thornton, A. Treves, L. Turin, R. Yuste, Q. Zhou. We also thank the following for other input: J.M. Alonso, M. Bethge, M. Deschenes, T. Elliott, A. Hyvarinen, W. Kim, M. Larkum, K. Likharev, K. Martin, A. Radulescu, T. Sejnowski, M. Sherman, A. Sillito, G. Stuart, H. Swadlow, R. Yuste, Q. Zhou