INSTITUTE FOR NEURAL COMPUTATION
University of California, San Diego
Institute for Neural Computation
La Jolla, CA 92093-0523
Technical Report #0401 18 March 2004 inc2.ucsd.edu

A Theory of Cerebral Cortex

Robert Hecht-Nielsen
Computational Neurobiology, Institute for Neural Computation, ECE Department,
University of California, San Diego, La Jolla, California 92093-0407 USA, rh-n@ucsd.edu

Abstract

A theory of the function of cerebral cortex and thalamus is sketched.

1. Introduction

This report sketches the author's theory of cerebral cortex and thalamus; incorporating relevant discoveries, improvements, changes and updates as of the issue date. The main focus is to provide a coherent, integrated picture of that portion of the theory which provides answers to the following questions: What is cortical knowledge and how is it acquired and stored?, and How is cortical knowledge used to carry out thinking? The theory's explanation for another key aspect of cortical and thalamic function – the moment-by-moment selection, evaluation, and execution of the *action commands* that control waking brain activity and muscle movement – is also briefly addressed.

This theory proposes the first cogent answers to these questions and therefore should be of interest to both scientists and technologists. In particular, the theory's characterizations of knowledge and thinking are radically different from the 'reasoning' approaches that others have considered; and, given their efficacy (as demonstrated herein via computer thinking experiments), these deserve attention on their own merits.

The reader is warned that this theory is not a simple 'sound bite' discovery; as is today's fashion. The pieces of this theory emerged gradually over decades (I began research in computational neurobiology in 1968). However, these pieces only came together as a whole late in 2003. Thus, this report must introduce and relate a number of novel concepts and show how they fit together. To keep the discussion accessible, detailed mathematical models of feature attractors and knowledge links are not discussed. The mathematics of confabulation is so simple that it is included.

The basic claim of the theory is that the cerebral cortex and thalamus implement thousands of *lexicons*, each containing thousands of *symbols*. These constitute the *terms of reference* used to describe the *objects* and *actions* of the mental universe. Each such lexicon is postulated to be implemented by a single, modular, thalamocortical neuronal circuit called a *feature attractor* (each symbol of the lexicon is represented by a distinctive population of co-active neurons). Each cortical *item of knowledge* (of which the average human has billions) is hypothesized to take the form of a unidirectional neuronal linkage between the neurons representing one symbol in one lexicon to neurons representing a second symbol in a second lexicon. Simultaneous inputs from (potentially) thousands of knowledge links to a feature attractor cause many of the neural ensembles representing symbols of its lexicon to become partially excited. The feature attractor then *operates*; extinguishing all of these excited neurons except those of the single most excited symbol ensemble. This simple, fast, 'winner-take-all' information processing operation, termed *confabulation*, is postulated to be the fundamental mechanism of vertebrate cognition.

The next section of this report (derived from a stand-alone article that has not yet been published) discusses an idealized mathematical description of confabulation. This discussion illustrates, largely sans biology, the essence of cortical knowledge acquisition, knowledge storage, and thinking. The third section then overviews how these mechanisms of knowledge storage and confabulation are implemented in biological tissue. This

third section also introduces the hypothesized process by which cortex controls waking brain function moment-by-moment. Finally, a brief discussion section ends the report.

While this theory is distinctly novel, many of its individual elements have been proposed long ago. A recent book chapter [Hecht-Nielsen 2003] details past research that influenced the early development of this theory.

2. Confabulation

Summary: The main research program of Artificial Intelligence has been to find some sort of reasoning process that can account for animal cognition. Confabulation, a symbolic prediction technique introduced here, is the antithesis of this program. Instead of trying to reason from assumed facts to conclusions, it uses a strange form of knowledge to rule out unreasonable outcomes: leaving a core set of not-unreasonable possibilities called an *expectation*. Then, employing this same knowledge in a different way, it identifies the highest quality conclusions in the expectation. Unlike reasoning, which must employ highly refined and specific knowledge that is difficult and expensive to obtain, confabulation uses a type of knowledge (conditional probabilities between pairs of symbols) that is simple and easy to obtain in the required vast quantities. The recently discovered neuronal *Marder / Turrigiano / Hebb* learning principle produces this kind of knowledge. Confabulation is proposed as the underlying mechanism of all vertebrate cognition.

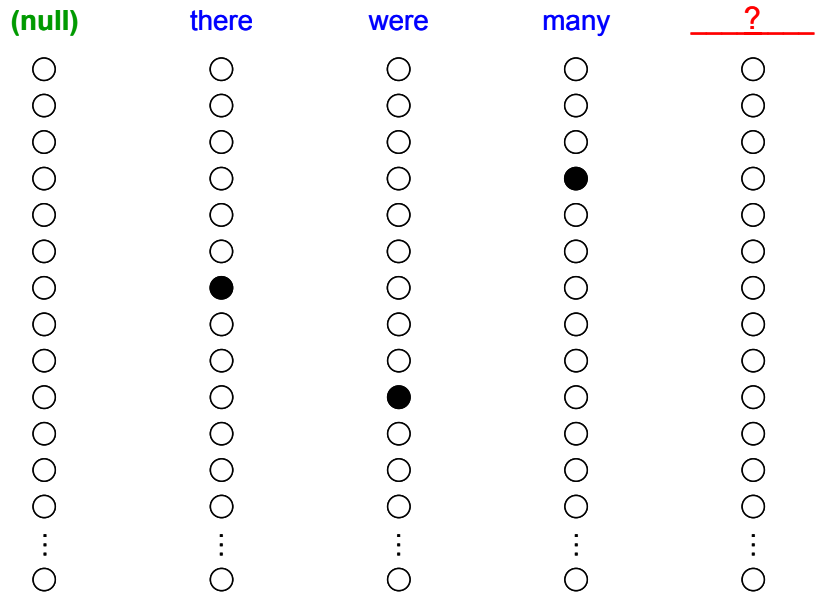

Figure 1. A confabulation architecture for answering phrase completion questions. Given a phrase (a string of one, two, three, or four words – in this case the phrase there were many) for which the set of reasonable next words (an *expectation*) are to be predicted, we place each word, in sequence, above its column of symbols (each represented by a circle) so that the last word of the phrase is above the fourth column from the left. Each column has 10,000 symbols, only a few of which are shown, representing the 10,000 word English lexicon we are using. The unique word assigned to each symbol is arbitrary, but is fixed at the outset and not allowed to change. For each word of a phrase we *activate* the symbol of its column that represents that word (the active symbols are the filled circles). Any leading columns which have no phrase words are left blank. At the end of the confabulation process the 'best' answer to the phrase completion question (the expectation symbol with the highest *estimated quality*) is expressed as an active symbol on the fifth column.

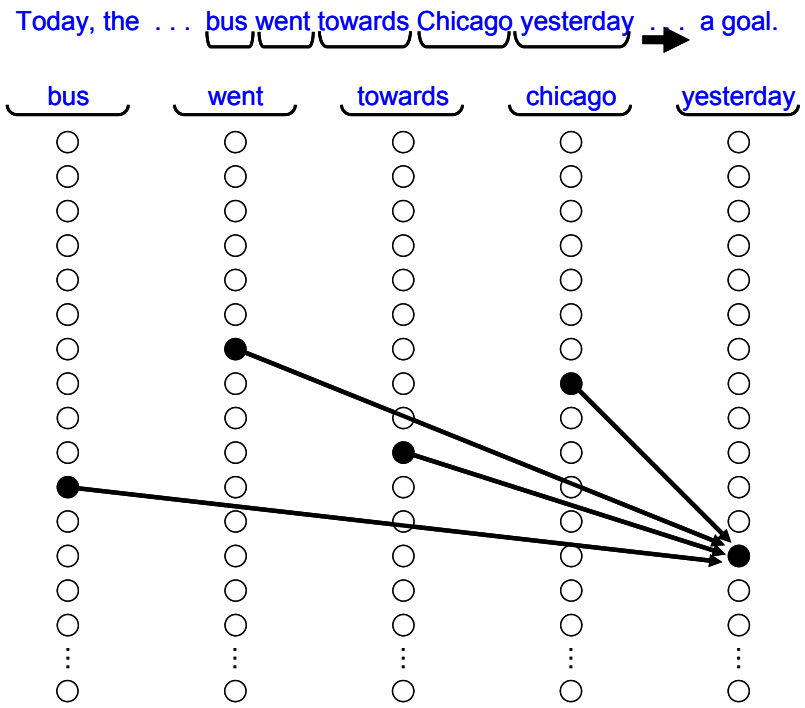

Figure 2. An example of the knowledge acquisition process. A five-word window (illustrated at top) marches, one word at a time, from the beginning of a 1.4 billion word English training corpus (the text string shown at top which begins with Today and ends with goal.) to the end. Each time an admissible substring (see text) is found in the window it is right-justified to the fifth column and expressed on the architecture by activating the appropriate symbols; as shown here for the admissible substring bus went towards chicago yesterday. For each such admissible substring a unidirectional weighted pairwise *link* is created from each active symbol on columns one, two, three, and four to the symbol active on column five (or, if the link already exists, its use count is incremented by one). All the links terminate on symbols of the fifth (*answer*) column, since that is where the knowledge is needed to select the reasonable phrase completion words. After the window has completed its traverse of the training corpus, links that are considered *meaningful* are retained and have a *weight* assigned to them. Each such link is an individual *item of knowledge*. 5,251,335 knowledge items were generated in this experiment.

Figures 1 and 2 illustrate a concrete example of confabulation. In this architecture, five *lexicons* (shown as columns of circles), each containing 10,000 *symbols*, are employed. In these example lexicons, each symbol is arbitrarily (and permanently) assigned to represent exactly one unique English word of the 10,000 most frequently seen words in a 1.4 billion word corpus of general proper English text (e.g., news stories). Concatenating the sentences of the corpus into one long string of words with punctuation (and eliminating capitals and replacing words outside the set of 10,000 with a special "not in vocabulary" marker), a five-consecutive-word window is marched from the beginning of this string to the end, advancing one word at a time (as illustrated in Figure 2). Each time a string of contiguous words (each among the 10,000) with no internal punctuation (period, comma, etc.) are found at the right end of the window, this complete *admissible substring* (which might be 2, 3, 4, or 5 words long) is used for *training* the architecture.

To carry out training, each word of the admissible substring is placed, in order, above the successive columns of the architecture; with the last word of the substring always placed above the fifth column (as shown in the example of Figure 2 for an admissible string of five words). In each column with a word above it, the unique symbol representing that word is *activated*. A pairwise unidirectional *link* (for which a *use count* is maintained) is then formed between each active symbol in the first four columns and the symbol active on the fifth column (or, if this link was already formed, its use count is incremented by one).

After the window has completed its long march down the training corpus string, each link that was formed and used at least three times (links with usage below this *cut-off* value are considered not meaningful and are

discarded) has a *weight* assigned to it, equal to the ratio of its use count and the number of times its *target symbol* appeared in admissible strings. This ratio approximates the conditional probability $p(\lambda_\theta|\varepsilon_5)$ of the appearance of the link's *source symbol* (the symbol of column θ having index λ_θ), given that its *target symbol* (the symbol of column five having index ε_5) is the symbol currently active on the fifth column (this link weight value is termed the *antecedent support probability level* of the link). Links having antecedent support probability below 0.001 are considered potentially unreliable and not meaningful and are discarded (in effect, deeming their correct $p(\lambda_\theta|\varepsilon_5)$ value to be 0). Each link is considered an individual *item of knowledge* and the set of all links from the symbols of one column to the symbols of the fifth are referred to collectively as a *knowledge base*. The reason why this particular type of probability is used, and not some other, is explained below.

Carrying out this knowledge acquisition process on a computer (with the above-mentioned English corpus) yielded 5,251,335 items of knowledge (the knowledge bases from columns one, two, three, and four to column five have 870,070, 1,293,451, 1,520,170, and 1,567,644 links, respectively). Now endowed with knowledge, the architecture of Figures 1 and 2 can be used (as described below) to *complete* arbitrary English phrases (symbol strings $\alpha_1\beta_2\gamma_3\delta_4$) from one to four words in length. A symbol index being zero means that there is no symbol active in that lexicon (e.g., $\alpha_1 = 0$ means that the input phrase is three or fewer words long).

CRITICAL MATERIAL

The probability $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$, a new quantity introduced here, can be interpreted as the degree (relative to other column five symbols) to which symbol ε_5 can serve as a *prediction* or *conclusion to be drawn* from a particular ordered set of symbols $\alpha_1\beta_2\gamma_3\delta_4$ (termed the *assumed facts*). A key advantage of using $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$, termed the *ideal quality* of ε_5, is that it does not depend on the frequency of appearance of ε_5, only on the degree of ε_5's 'logical support' of, or 'level of consistency' with, the set of assumed facts $\alpha_1\beta_2\gamma_3\delta_4$. Thus, while finding the ε_5 which maximises this probability is not reasoning, it clearly has similarities to the reasoning process of proving or inferring ε_5 from $\alpha_1\beta_2\gamma_3\delta_4$. The set of all ε_5, if any, such that $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5) > 0$ is termed the *ideal expectation*. Of course, probabilities such as $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$ are conceptual and not practically calculable; but represent the benefit that ideal *omniscient knowledge* would confer.

Confabulation (defined here, for simplicity, only for the concrete case of exactly four assumed facts $\alpha_1\beta_2\gamma_3\delta_4$ with positive indices, but with obvious generalization) is a predictive technique that assumes possession of *all* non-zero (i.e., meaningful) antecedent support probabilities of the form $p(\lambda_\theta|\varepsilon_5)$ (and assumes that all other such probabilities are zero). Using this *exhaustive* knowledge, given assumed facts α_1, β_2, γ_3, and δ_4, confabulation creates the *expectation* consisting of all symbols ε_5 having $p(\alpha_1|\varepsilon_5)$, $p(\beta_2|\varepsilon_5)$, $p(\gamma_3|\varepsilon_5)$, and $p(\delta_4|\varepsilon_5)$ all positive. If the expectation is empty, then there are no reasonable predictions (knowing when we don't know is itself useful). Notice that if any of these pairwise conditionals are zero, say $p(\beta_2|\varepsilon_5) = 0$, then applying the identity $p(abcde) = p(a|bcde) \cdot p(b|cde) \cdot p(c|de) \cdot p(d|e) \cdot p(e)$ [Pearl] yields $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5) \equiv p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\varepsilon_5) = p(\alpha_1\delta_4\gamma_3\beta_2\varepsilon_5)/p(\varepsilon_5) = p(\alpha_1|\delta_4\gamma_3\beta_2\varepsilon_5) \cdot p(\delta_4|\gamma_3\beta_2\varepsilon_5) \cdot p(\gamma_3|\beta_2\varepsilon_5) \cdot p(\beta_2|\varepsilon_5) = 0$ (since an item of knowledge from β_2 to ε_5 does not exist). This proves the:

Expectation Theorem: Confabulation produces an expectation which contains the ideal expectation as a subset.

Notice in the above that we could just as well have used the probability $p(\beta_2\varepsilon_5)$ (which essentially corresponds to classical 'Hebbian' co-occurrence learning) instead of $p(\beta_2|\varepsilon_5)$; because, by Bayes' law, $p(\beta_2|\varepsilon_5) \equiv p(\beta_2\varepsilon_5)/p(\varepsilon_5)$. Below it is shown why $p(\lambda_\theta|\varepsilon_5)$ knowledge is the correct choice.

In some applications (e.g., solving the cocktail party problem [Sagi, et al.]), expectations are all that are needed. However, in other situations (e.g., for the phrase completion problem example considered here), we need, if possible, to find those ideal expectation elements of highest ideal quality (these are termed *ideal answers*). In lieu of ideal qualities $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$, which are unknowable, confabulation assigns an *estimated quality* to each expectation element ε_5 equal to the product $p(\alpha_1|\varepsilon_5) \cdot p(\beta_2\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)$. To see why this

makes sense, note that by the above identity we can write $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5) = p(\alpha_1|\beta_2\gamma_3\delta_4\varepsilon_5) \cdot p(\beta_2|\gamma_3\delta_4\varepsilon_5) \cdot p(\gamma_3|\delta_4\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)$. Similar expansions can be derived for the quantities $p(\beta_2\gamma_3\delta_4\alpha_1|\varepsilon_5)$, $p(\gamma_3\delta_4\alpha_1\beta_2|\varepsilon_5)$, and $p(\delta_4\alpha_1\beta_2\gamma_3|\varepsilon_5)$, which are all equal to $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$. Multiplying these four expansions together and simplifying (see APPENDIX for details) yields:

$$
\begin{aligned}
[p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)]^4 = {} & [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\alpha_1\varepsilon_5)] \\
& \cdot [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\beta_2\varepsilon_5)] \\
& \cdot [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\gamma_3\varepsilon_5)] \\
& \cdot [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\delta_4\varepsilon_5)] \\
& \cdot [p(\alpha_1|\varepsilon_5) \cdot p(\beta_2|\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)].
\end{aligned}
$$

The key observation is that, for natural information environments, the ratios in the first four bracketed quantities on the right side will, in general, each probably change only slightly for different expectation elements ε_5. For example, for text, consider $\alpha_1\beta_2\gamma_3\delta_4$ = the train was going south. Then, $p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\alpha_1\varepsilon_5)$ = p(the train was going south)/p(the __ __ __ south), $p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\beta_2\varepsilon_5)$ = p(the train was going south)/p(__ train __ __ south), $p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\gamma_3\varepsilon_5)$ = p(the train was going south)/p(__ __ was __ south), $p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\delta_4\varepsilon_5)$ = p(the train was going south)/p(__ __ __ going south). If south were replaced by any other expectation element (north, east, west, fast, slow, etc.) these ratios would probably change very little. Similarly, if $\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5$ were a movement (i.e., postural goal) sequence (e.g., an overhead arm swinging motion as used in swimming), changing ε_5 (an expected next postural goal after $\alpha_1\beta_2\gamma_3\delta_4$, e.g., from a downward continuation of the movement for the breast stroke to an outward movement for the butterfly), would probably not affect these probability ratios much; and so forth for visual scene descriptor symbols, etc. If the values of these four bracketed quantities do not change with different expectation elements ε_5, then the ordering of the ideal expectation elements will be the same for both ideal quality and for estimated quality, since the fourth power of the former (a direct function) will be a positive constant times the latter. Thus, the following conjecture is offered:

Estimated Quality Conjecture: For assumed facts in vertebrate cognitive information environments (sensory percepts, language, movement processes, thought processes, event memory sequences, abstract concepts, etc.), the ordering of ideal expectation elements by estimated quality and by ideal quality will generally be the same.

Typical answers obtained by applying confabulation to predict phrase completions using the English text knowledge bases described above are listed below [The answers, if any, obtained by applying estimated quality to expectation elements, are shown in descending order of estimated quality. When the expectation has more than six elements the top six answers are shown in parentheses, followed by the total number of conclusions in the expectation; if the expectation has six or fewer elements, square brackets are used, and all the answers are shown]:

- she could determine (whether, exactly, if, why, how, precisely) 8
- academic administrators said [faculty]
- if it was not (immediately, clear, enough, true, properly, stupid) >999
- earthquake activity was [centered]
- for lack of a (unified, blockbuster, comprehensive, definitive, coordinated, protein) 111
- a lack of (urgency, oxygen, understanding, confidence, communication, enthusiasm) 407
- regardless of expected [outcome, length]
- the bells were [spun, sounded, chanted]
- cars drove down a (lane, freeway, highway, dirt, taxi, tying) 9
- driving west on interstate [highway, freeway]
- snow fell in (freezing, montana, portions, northwestern, northeastern) 11
- the facts point to [ ]
- threats of terrorist [attacks, retaliation, strikes, violence]
- tomorrow will bring high [ ]
- the machine (tools, tool, guns, gun, operator, shop) 33

- babies can learn [ ]
- children can learn [lessons, math, english]
- students can learn [lessons, math, english]
- college students can learn [math]
- knowledge of historical [facts, subjects, styles]
- questions that cannot be (answered, solved, resolved, avoided, addressed, yes) 9
- flights delayed [ ]
- benefits from additional (cost-cutting, taxable, protections, taxes, acquisitions, payroll) 11
- limitations [expired, expires, imposed]
- dangerous areas [ ]
- her responsibility for taking [sole, matters]
- her responsibility for making [errors, matters, sure, references, choices, lethal]
- his responsibility for making (mistakes, matters, bombs, references, decisions, sure) 11
- his responsibility for taking [actions, sole, matters, decisions]
- mechanical failure [caused]
- bridges under construction [ ]
- the meaning of (adventure, one's, symbols, purpose, life, christmas) 15
- not only the facts (concerning, surrounding, regarding, relating, speak, underlying) 9
- the focus was on (speculation, pharmaceuticals, core, laser, stopping, improving) 156
- second only to the (dalai, ncaa, sonics, bathroom, podium, majors) 988
- an exclusive interview [ ]
- lead in the right (leg, path, wings, kidney, conclusion, wheels) 10
- a general (manager, motors, obligation, instrument, accounting, contractor) 529
- massive amounts of [radiation, fluid, copper, muscle]
- surplus (widened, narrowed, projections, plutonium, totaled, commodities) 45
- hearings on (capitol, whitewater, waco, brown's, allegations, questionable) 18
- crystal clear [glasses]
- how to clean up (toxic, contaminated, emissions, radioactive, pollution, debris) 16
- bottom of the (ninth, eighth, pan, seventh, baltic, closet) 111
- pockets of (resistance, poverty, weakness, dissent, infection, corrupt) 10
- a reasonable chance that [succeeding]
- the cancer caused [spreads, lung, kidney, birth, chronic, infections]
- lack of a better (understanding, coordination, description, appreciation, representation, method) 30
- crime rates (soared, soaring, plummeted, fueled, surged, dropping) 9
- the end (zone, optional, here, of, user, shannon) 227
- it was a (reminder, mistake, nice, joke, logical, surprising) >999
- crowded (commuter, marketplace, subway, courtroom, skies, sidewalk) 62
- they crowded (onto, lobby, shopping, shelters, around, into) 22
- beaches are covered with [pools]
- there were many (indications, surprises, instances, casualties, signs, exceptions) >999
- are easy to (install, dismiss, detect, locate, accumulate, criticize) 159
- microsoft makes software for [apple's, desktop, hardware]
- the green car turned [yellow]

A more elaborate experiment, utilizing 100,000 symbols in each lexicon (representing 30,000 words and 70,000 word groups) trained on a similar text corpus, but with knowledge bases linking every ordered pair of lexicons (i.e., this system can fill in a missing word anywhere in a phrase), yielded:

- caused by (chlorine, emission, concern over, smog, bacterial) 17 pollution
- sail the ship to (arrive at, papeete, international waters, antarctica, mururoa, auckland) 8
- bottom line (was boosted, also got a boost, can pass, here is, got a boost, by firing) 23
- alternate sorts of [ ]
- later (made his way, strolled, while driving, marched, tore, dragged) 63 down the street
- overcame the advantages of [ ]

- a belt (was wearing, tucked into, strapped to, hanging from, he grabbed, he wears) 72
- the belt (he was wearing, buckle, in malta, sander, loops, after failing) 40
- taxes must (live within, cut spending, cut expenses, spend more, adhere, stimulate growth) 67
- quarterback scored the winning [touchdown, td]
- some [trainees, buddhists, winters, cadets, geese, choreography] can be taught to
- not reached the age of [consent, parity, adulthood]
- impending (dividend payments, quarterly dividend, plant closings, second trial, republican presidential nominee, retrenchment) 29 will be
- (vichy, gen. noriega's, south africa's apartheid, totalitarian, ayatollah khomeini's, dictatorial) 537 regime
- practicing for intense [air show]
- the huge waves (whipped, crashing, tore, ripped, smashed, buffeted) 8
- destructive forces of [disrupting]
- breach in security (cordon, embankment, duties, obligations, safeguards, obligations under) 10
- doctors purpose was (prescribing, to inform, harming, assisting, evaluating, persuading) 8
- overcoming past [objections to, prejudices, economic problems, rivalries, reluctance, tyranny]
- discovering the key (enzyme, brain cells, link between, flaw, proteins, culprit) 26
- the process could (eventually lead, proceed without, be completed within, squeeze out, fall apart, gain momentum) 237
- the constituents [he served, ordinarily, twisting]
- under the christmas (tree, lights, cold snap, trees, furlough, carol) 25
- higher than the (first quarter's, national average, consensus forecast, previous record, year-earlier level, previous night) 173
- wandered along the [fringes, aisles, wooded, stacks, desert, shadowy]
- the train was (so crowded, still burning, headed south, scheduled to arrive, headed north, thrown off) 203
- demonstrators chanted [slogans, anti-american, anti-government, anti-nuclear, loudly]
- demonstrated her mathematical [prowess]
- the rules determined [exactly how much, how fast, what kind, how much, how many, diplomacy]
- dealing with advanced [concepts, old age, shooters]
- reached its maximum (altitude, height, legal limit, time limit, limit, three-year) 7
- (hardware and, web-based, provider of, netscape's, unix, company called) 21 software solutions
- (mobile, sonic, netscape, advances in, carlton, ntt) 140 communications were made
- was overcome by (thick smoke, smoke from, emotion, tear gas, smoke, fumes) 11
- mickey and minnie [mouse]
- morning brought (relative calm, clear skies, scattered showers, a pledge, heavy rain, high winds) 26
- provided a (fairness option, $## million cash infusion, glimpse into, detailed accounting, full accounting, major boost) 686
- gave way to the (realization that, realities of, southern edge, skies, sunshine, exuberance) 15
- i (glanced, stared, peered, used to live, strolled, gazed) 158 down the street
- later (made his way, strolled, while driving, marched, tore, dragged) 63 down the street

This latter architecture possessed 63,871,367 items of knowledge. Notice the automatic and instantaneous emergence of 'grammar' and 'semantics' in the examples of both experiments. Confabulation likely obviates the need for linguistic concepts such as grammar and semantics; and much else.

The name *confabulation* derives from the psychiatric syndrome of confabulation, in which the production of language is not impaired, but key facts regarding the context of a production seem to be unavailable. The results of thought in this case are conclusions (often not factually correct) that cannot be ruled out based upon the reduced set of assumed facts employed. The confabulation syndrome thus reveals the inner working of thought as the identification of not-known-to-be-wrong conclusions; as opposed to some sort of reasoning. Thus the name.

One of the most important aspects of confabulation is its ability to effectively deal with novel presentations of familiar elements (a characteristic that today's information processing approaches generally lack). Some of the

examples in the above experiments are novel phrases that were never encountered in the training corpus. Since only pairwise antecedent support knowledge is used, the system has no ability to decide that a particular collection of assumed facts is novel or familiar and deals effectively with both; as long as they 'make sense' (i.e., do not have any zero probabilities within the formula relating estimated to ideal quality). Obviously, as evidenced by the inability of confabulation to complete some sensible example phrases in the above experiments (e.g., bridges under construction [ ]), the knowledge bases used are not exhaustive. Nonetheless, every meaningless test phrase we have tried (e.g., "tune card fly bold") yields an empty expectation. Thus, somewhat incomplete (i.e., not completely exhaustive) knowledge seems to translate into a tendency to make errors of omission, not commission; a generally favorable attribute.

Note that in the above formula relating estimated quality to ideal quality the 'Hebbian' $p(\lambda_\theta\varepsilon_5)$ quantities would not work: we must have the $p(\lambda_\theta|\varepsilon_5)$. To acquire one item of this knowledge, symbols λ_θ and ε_5 must co-occur and a unidirectional link must be formed from λ_θ to ε_5 having strength $p(\lambda_\theta|\varepsilon_5)$ (or some monotonically increasing function of this quantity; such as the logarithm of the ratio of $p(\lambda_\theta|\varepsilon_5)$ to p_0, the minimum meaningful link strength). In brains, the phenomenon of long term potentiation [Cowen, Sudhof, and Stevens] (and the suspected associated, although as yet undiscovered, permanent memory solidification process) supports the notion of Hebb that causal co-occurrences of firing of λ_θ and ε_5 will lead to a synaptic link of strength equal to a direct function of the joint co-occurrence probability $p(\lambda_\theta\varepsilon_5)$. Recent studies of post-synaptic neurotransmitter depolarization transduction response by Marder and her collegues [Marder] and by Turrigiano and her colleagues [Turrigiano; Desai] suggests that the post-synaptic apparatus of an excitatory cortical synapse is independently modifiable in multiplicative series with this $p(\lambda_\theta\varepsilon_5)$ efficacy; tending towards a neurotransmitter receptivity proportional to a direct function of the reciprocal of its neuron's average firing rate; which is essentially $p(\varepsilon_5)$. The net result is implementation by this *Marder / Turrigiano / Hebb* learning process of an overall link strength directly related to $p(\lambda_\theta\varepsilon_5)/p(\varepsilon_5)$, which, as noted above, is $p(\lambda_\theta|\varepsilon_5)$. Thus, it is plausible that biological learning processes at the neuron level can accumulate the knowledge needed for confabulation. More on the biological mechanization of these knowledge links below.

Even if exhaustive knowledge is available, all the expectation theorem guarantees is that no viable answers (i.e., elements of the ideal expectation) will be left out of the confabulation-derived expectation. So, the potential exists for an expectation to also contain many 'spurious' symbols (unreasonable predictions) that just happen to have all their $p(\lambda_\theta|\varepsilon_5)$ probabilities positive. Further, even if the estimated quality ordering of ideal expectation elements is the same as that for ideal quality, there is no guarantee that when estimated quality is applied to such spurious symbols that these estimated quality values will not exceed those of the ideal answers. [Additionally, some seemingly straightforward predictions by confabulation might, in fact, not be correct because a zero higher-order probability value (indicating an unusual deviation of this case from 'normal' patterns of usage) would invalidate the estimated quality to ideal quality conversion formula. However, these cases may not actually be a problem because they can be handled by explicitly and exhaustively learning all of the involved *exceptions*. For example, 'an historic' as opposed to 'a historic.'] Obviously, it would be easy to design contrived symbolic information environments that strongly exhibit the above problems, and others. As a result, confabulation might seem disturbingly inadequate and possibly even dangerous as a basis for cognition.

The unseen hand of evolution is probably why confabulation works so well in practice. Text is an ideal example of this: an information environment presumably explicitly designed to exactly fit the capabilities and function of confabulation; which would explain why none of the possible problems mentioned in the last paragraph are seen in the text experiments presented above. It seems likely that the symbolic lexicons of today's vertebrates arose by a long and arduous process of evolution driven by the survival benefits conferred by having expectations with no spurious elements and estimated quality provided by confabulation closely approximating that of the ideal. Once achieved, these lexicon designs have been highly conserved. In other words, it is conjectured that proper design of the lexicons, and proper design of the learning saga that each individual undergoes during development – both genetically determinable – can control deviations of confabulation from the ideal.

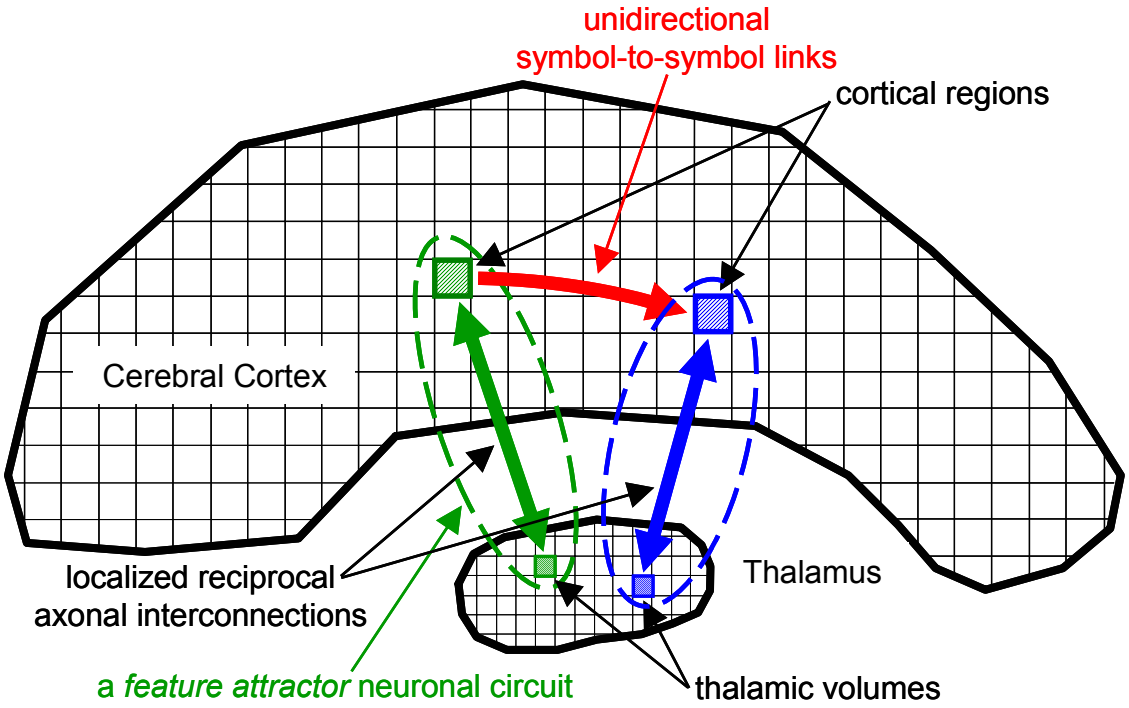

Figure 3 Corticothalamic information processing functionality. The cortical theory holds that the quarter-square-meter surface of human cerebral cortex is effectively divided into thousands of small localized notional patches or *regions*. Each such cortical region is hypothesized to be paired with a unique, much smaller, dedicated *volume* of thalamus; to which it is reciprocally axonally connected. Each such pair of a cortical region and its thalamic volume, along with their reciprocal connections, constitutes a distinct modular *feature attractor* neuronal circuit. A feature attractor circuit, when allowed to run free, is hypothesized to function as an 'analog to symbolic' converter: rapidly converting an arbitrary ensemble of neuronal excitations on its cortical region into one, and only one, of a fixed *lexicon* of sparse neuronal activity patterns on that same cortical region (each such pattern consisting of a few hundred highly active neurons, representing one *symbol* of the lexicon). These regional lexicons are formed at various points in childhood and then largely frozen for life. Each lexicon describes exactly one *attribute* of an *object* or *action* of the mental world. These lexicons provide the fixed symbolic terms of reference that are required for the accumulation and use of knowledge over decades. Pairs of symbols which appear together meaningfully have unidirectional neuronal links established between them via 'Hebbian' synaptic modification. Each such link is a single *item of knowledge*; of which the average human is postulated to possess billions.

Clearly, confabulation immediately and automatically applies all available knowledge concerning each assumed fact (links from that assumed fact's symbol to the symbols of the answer lexicon) that would bear on the selection of the answer to a question. Thus, confabulation is immune to the *knowledge selection problem* (somehow deciding which items of knowledge to apply to which arguments at each step of information processing); another difficulty which bedevils knowledge-based information processing schemes.

The time required to carry out confabulation is that required to create the expectation, order its symbols by estimated quality, and select the top item. As explained in more detail later, the cerebral cortex is hypothesized to be able to carry out this process very rapidly using localized, parallel processing. Even though neurons have very low speed and dynamic range, a complete confabulation operation can probably often be completed in a few tens of milliseconds (ms). The theory holds that confabulation (and its variants) is the only elemental information processing operation of vertebrate cognition. It is the core essence of thinking. For confabulation to be the basis of thinking requires that every *attribute* used to describe an *object* or *action* of the mental world must be represented by a single symbol drawn from a lexicon dedicated to that attribute. In other words, a 'column' of symbols needs to be provided for each of a multitude of individual descriptive attributes. Further, once formed, each such column must not have symbols deleted from it, or changed in meaning, for life

(addition of symbols can occur); in order to ensure that accumulated items of knowledge which employ those terms of reference can be used at any point in the future. That is precisely what the theory hypothesizes.

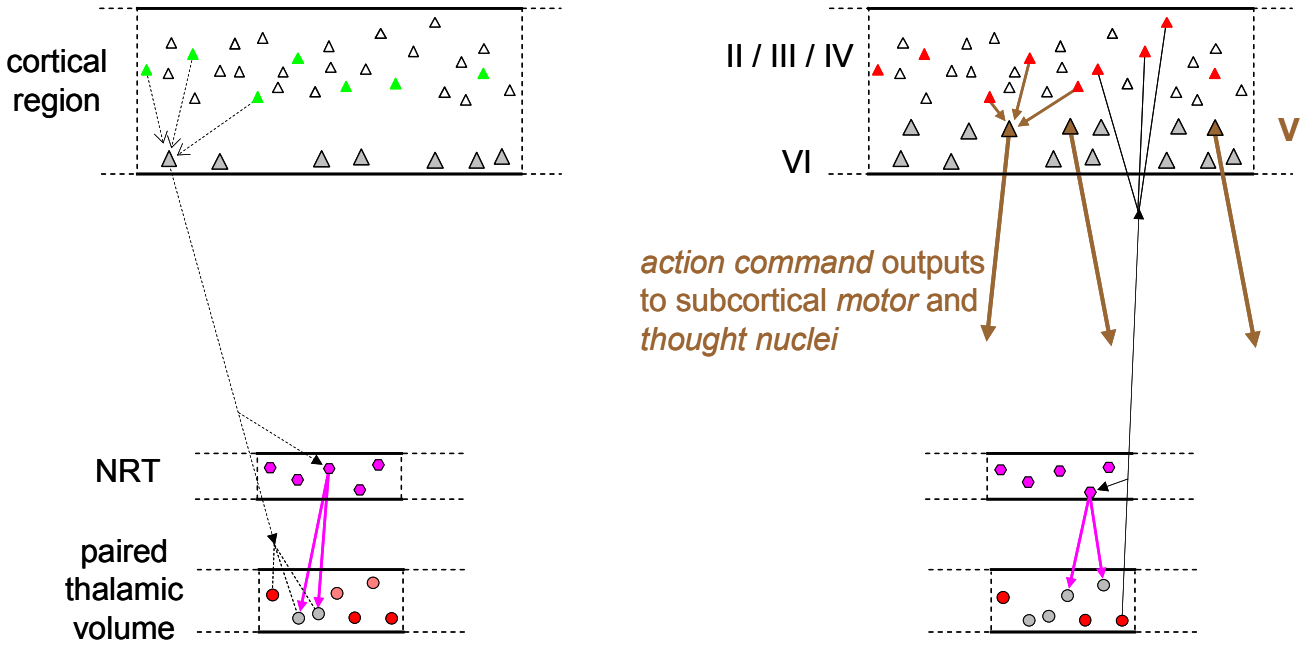

Figure 4 Overview of the normal function of a single mature feature attractor neuronal circuit. The green neurons of the cortical region (shown in the left diagram, which shows the first, downward transmission, step of operation of the circuit) are a few of (potentially) thousands of pyramidal *feature detector* neurons of cortical layers II, III, and IV that have been excited by external thalamic and/or cortical inputs to this cortical region. These excited feature detector neurons together send signals to neurons of layer VI of the region; a few tens, to many hundreds, of which thereby become excited. These excited layer VI neurons, in turn, send outputs to a collection of glomeruli in the paired thalamic volume of the feature attractor circuit; causing a few tens to hundreds of them to become excited. The right diagram shows these excited glomeruli sending their output back up to layer IV of the cortical region of the circuit. Here, only neurons representing exactly one of the lexicon symbols (the one that most closely matches the initial ensemble of excited neurons) become activated. Each lexicon symbol is represented by a fixed-for-life set of hundreds of feature detector neurons, tens of layer VI neurons, and tens of glomeruli (because these sets are selected essentially randomly they automatically have low maximum pairwise overlap). The surprising fact that a simple model of this circuit, starting with multiple symbols initially partially excited, is guaranteed to eventually converge to a single symbol after a finite number of round trips (for at least certain initial states) was established mathematically in 1988 by the author and Karen Haines [Haines and Hecht-Nielsen; Sommer and Palm; Kosko; Hopfield; Amari]. The added surprise that this convergence actually occurs in a single round trip was discovered experimentally by the author in the early 1990s (when the first extensive computer experiments with models of feature attractors at sufficiently large scales first became feasible). The even greater surprise that this circuit will often converge to a single symbol in a single round trip for an essentially arbitrary initial state was only established recently by the author [to be published]. Note that both the downward and the upward connections also target the thalamic reticular nucleus (*NRT*) (which then has an inhibitory effect on selected glomeruli of the thalamic volume). For simplicity, the NRT is ignored here.

Many cortical thought processes only require expectations, not answers (e.g., for operations such as *segmenting* an attended object from a complex dynamic, multi-object sensory input stream so that it can be analyzed in isolation – an example is the *cocktail party problem* [Sagi] where only the feasible sound inputs from one speaker among many are allowed to proceed past the primary auditory cortical level).

The next section sketches the theory's explanation of how knowledge is stored and how confabulation is implemented in the cerebral cortex. The separate, but strongly related, question of how *thought processes*

(stored coordinated sets of confabulation operations) and *movement processes* (stored coordinated sets of muscle operations) are *selected* and *executed* (in this theory thoughts and movements are viewed as siblings and are referred to – both individually and in combinations – as *actions*) at each moment during wakefulness is also briefly considered.

3. Thalamocortical Information Processing

For concreteness, unless otherwise indicated, this discussion will center on mature human cerebral cortex and thalamus. This report is only intended to provide an introductory sketch of the still-developing theory (primarily for use as background for more focused research reports). For this purpose, many details are suppressed and other topics that the theory comments upon (e.g., how confabulation is used to carry out sensory and motor processing) are only briefly mentioned, or are skipped altogether. A previous description [Hecht-Nielsen 2003] of portions of the theory provides further elaboration of some of its biological aspects. Familiarity with general human neuroanatomy and neurophysiology is assumed.

As shown in Figure 3, the theory hypothesizes that cerebral cortex (along with thalamus) implements confabulation using a fundamental modular neuronal circuit called a *feature attractor*. Each of these circuits implements exactly one lexicon of symbols and carries out the confabulation processing operation when that lexicon is used as the answer lexicon of a confabulation operation. On other occasions, when such a lexicon is used to express an assumed fact symbol, its feature attractor circuit assumes that pure symbolic state and holds it active for the brief time needed for the answer lexicon(s) to receive the link inputs from that symbol and complete the confabulation operation. Just as with movements, precise relative timing of the individual events involved in carrying out confabulation is essential. This is because a feature attractor circuit that is to carry out confabulation must have all the necessary link inputs (often thousands of them, emanating from sometimes more than ten assumed fact symbols) arrive at the cortical region of the circuit at precisely the same time. Further, that 'answer lexicon' circuit must usually be in a 'cleared' or 'erased' state at that moment. The signal that then triggers confabulation (actually, this is the suspension of a signal – as described below) must arrive a very short time after these input link signals have arrived and excited tens of thousands of neurons within the circuit. The items of knowledge used in thought are stored in cortex as unidirectional physical links connecting the neurons of a symbol in one feature attractor's lexicon with a symbol in another's. Knowledge bases of these links (where, again, a *knowledge base* is the set of all links in one direction between symbols of a particular pair of lexicons) must themselves be deliberately and individually enabled by control signals at the proper moment. The details of these postulated biological implementations are now described.

3.1 Feature Attractor Neuronal Circuits: The Engines of Thought

To keep the exposition simple, feature attractors will be treated as isolated, independently operated, modular units that do not functionally interact with one another except via knowledge links being used to convey assumed fact link probabilities. In reality, the picture is more complicated: collections of feature attractor circuits having nearby cortical regions often do interact somewhat (for example, they can cooperate in their choices of symbols to keep overall information redundancy under control) and are often operated in synchrony. However, for the purposes of this report's 'first-order' sketch of the theory, these refinements can be ignored.

As shown in Figure 4, each modular feature attractor neuronal circuit has a cortical *region* of localized extent with a cortical surface area of a few square millimetres (perhaps 2-10 mm^2). This cortical region contains hundreds of thousands of neurons, of which perhaps more than 100,000 are (excitatory, pyramidal) *feature detector* neurons available to function as representations for the symbols of its lexicon (each lexicon typically contains a few thousand symbols). By a process akin to vector quantizer (VQ) codebook development in information theory [Zador] the symbols of a lexicon develop together over a period ranging from weeks to years, typically during childhood. Each symbol is represented by a specific set of a few hundred feature detector neurons. Because of the random selections involved, the maximum neuron overlap between any pair of such sets will, with extremely high probability, be relatively small.

The symbols of each individual lexicon are developed so as to represent one specific attribute of an object or action with low information loss. The attribute a particular lexicon describes is determined genetically by the

nature of the sources of the axonal inputs (from thalamus, cortex, or elsewhere) that happen to land on that particular region. Human cerebral cortex / thalamus implement thousands of these modular feature attractor circuits; one for each lexicon. The process by which each lexicon is developed is now sketched.

In the beginning, the feature attractor circuit of the lexicon is immature and not functional. First, a sizable subset of the roughly 100,000 feature detector neurons of cortical layers II, III, and IV of the region happen by chance to preferentially receive external inputs and are stimulated repeatedly by these inputs. These neurons develop, through various mutually competitive and cooperative interactions, responses which collectively cover the range of signal ensembles the region's input channels are providing. In effect, each such feature detector neuron is simultaneously driven to respond strongly to one of the input signal ensembles it happens to repeatedly receive; while at the same time, through competition between feature detector neurons within the region, it is discouraged from becoming tuned to the same ensemble of inputs as other feature detector neurons of the region. This is the classic insight that arose originally in connection with the mathematical concept of *k-means*. These competitive and cooperative VQ feature set development ideas have been extensively studied in various forms by many researchers from the 1960s through today, including Queen [Nilsson], [Tsypkin], [von der Malsburg], [Grossberg], [Carpenter], and [Kohonen]. The net result of this first stage of feature attractor circuit development is a large set of feature detector neurons (which, after this brief initial plastic period, become largely frozen in their responses – unless severe trauma later in life causes recapitulation of this early development phase) that have responses with moderate local redundancy and high input range coverage (i.e., low information loss). These might be called the *simple* feature detector neurons. They probably constitute roughly half of the total available population of the region's feature detector neurons.

Once the simple feature detector neurons have been formed and frozen, additional *secondary* (or "complex") feature detector neurons within the region then organize. These are neurons which just happen (the wiring of cortex is locally random and is essentially formed first, during early organization and learning, and then is soon frozen for life) to receive most of their input from simple feature detector neurons (as opposed to primarily from extra-regional inputs, as with the simple feature detector neurons themselves). This often implies that the simple neurons which feed them will tend to be rather more widely dispersed than the external inputs, particularly in the case of sensory regions where the external simple cell inputs are being supplied primarily by thalamic axons projecting to cortical layer IV; which have a very limited distribution range from the point where they enter cortex. These new secondary feature detector neurons also organize along the lines of a VQ codebook – except that this codebook sits to some degree 'on top' of the simple cell codebook. The net result is that secondary feature neurons tend to learn statistically common combinations of multiple co-active simple feature detector neurons.

A new key principle postulated by the theory relative to these populations of feature detector neurons, is that secondary feature detector neurons also develop inhibitory connections (via growth of axons of properly interposed inhibitory interneurons that receive input from the secondary feature detector neurons) that target the simple feature detector neurons which feed them. Thus, when a secondary feature detector neuron becomes highly excited (partly) by simple feature detector neuron inputs, it then immediately shuts off these simple neurons. This is the theory's *precedence principle*. In effect, it causes groups of inputs that are statistically 'coherent' to be re-represented as a whole ensemble; rather than as a collection of 'unassembled' pieces.

Once the secondary feature detectors have stabilized they too are then frozen and tertiary feature detectors (often coding even larger complexes of statistically meaningful inputs) form their codebook. They too obey the precedence principle. For example, in primary visual cortical regions, there are probably tertiary feature detectors which code long line segments (probably both curved and straight) spanning multiple regions. This is one example of how nearby regions might interact – such tertiary feature detectors might well inhibit and shut off lower-level feature detector neurons in other nearby regions. Of course, other inhibitory interactions also develop – such as the line 'end stopping' that inhibits reactions of line continuation feature detectors beyond its end. In essence, the interactions within cortex during the short time span of its reaction to external input (a few tens of milliseconds) are envisioned by this theory as similar to those postulated by Stephen Grossberg and Gail Carpenter and their colleagues in their visual processing theories [Grossberg 1987, Carpenter and Grossberg 1991].

Once the feature detector neurons (of all orders) have had their responses frozen, the next step is to consider the sets of feature detector neurons which become highly excited together across the cortical region due to external inputs. As mentioned above, because the input wiring of the feature detector neurons is random; the feature detector neurons function somewhat like VQ codebook vectors with many of their components randomly zeroed out (i.e., like ordinary VQ codebook vectors projected into randomly selected low-dimensional subspaces). In general, under these circumstances, it can be established that any input to the region (again, whether from thalamus, from other cortical regions, or from other extracortical sources) will cause a roughly equal number of feature detector neurons to become highly excited. This is easy to see for an ordinary VQ codebook. Imagine a probability density function in a high-dimensional input space (the raw input to the region). The feature detector responses can be represented as points spread out in a roughly equiprobable manner within this data cloud (at least before projection into their low-dimensional subspaces). Thus, given any specific input, a roughly uniform number of highest appropriate precedence feature detector points will be close to that input vector – thus causing their corresponding neurons to become highly excited. Thus, because of the information theoretic properties of VQs, the histogram of feature detector responses to each external input will tend to be roughly the same for all inputs. The theory postulates that each such ensemble of excited feature detector neurons possesses a few hundred members.

What happens next is that, as external inputs continue to arrive at the feature attractor circuit's cortical region, the thalamocortical interaction portion of the circuit begins to organize. The idea is that the circuit has the propensity for an ensemble of simultaneously excited feature detector neurons of its cortical region to form connections to pyramidal neurons of layer VI of the cortical region. These connections are hypothesized to seek out neurons which happen to send excitation to thalamic glomeruli which just happen to send excitation back to the currently excited cortical neurons (either directly or through precedence principle predecessors). How this works, I am not sure. But the general idea is that, before long, many excited feature detector neuron ensembles become directly axonally connected to a paired, but much smaller, ensemble of layer VI neurons (perhaps containing a few tens of neurons). This Layer VI ensemble is then directly axonally connected to a small ensemble of thalamic glomeruli (again, a few tens of glomeruli). This ensemble of glomeruli then becomes directly axonally connected back to the original feature detector ensemble (or, at least, to its lower-precedence neurons).

Once these loops form, when they function, they tend to perseverate for a few tens of milliseconds (freezing the excitement states of the neurons of the feature attractor which are participating for this time). During this brief period, these axonal connections tend to become strengthened and partially frozen. As a result, discrete triples of ensembles of feature detector neurons, Layer VI neurons, and thalamic glomeruli begin to form. Thousands of these triples arise as the circuit is exposed to a wide sampling of its input. However, after a while, so many of these triples form (covering the full range of input possibilities) that each new input tends to simply activate one of the existing, strengthened, triples; rather than creating a new triple. In other words, the circuit becomes possessed of thousands of fixed stable states; into the nearest matching of which the circuit will fall when presented with a new input (after the brief period of self-interaction that occurs immediately after a new input is received).

That this thalamocortical circuit should exhibit such a remarkably simple and uncomplicated dynamical 'attractor' behavior is almost amazing. Extensive past studies of attractor networks [Amari 1972, 1974; Hopfield; Cohen; Kosko; Amit; Sommer and Palm] have never described a network with such desirable dynamical characteristics. Quite the contrary; most attractor networks have been shown to be subject to many serious dynamical shortcomings, such as unwanted stable states, poor or no convergence to desired stable states, slow convergence, etc. As discussed briefly in the caption of Figure 4, a simple mathematical model that I have constructed of the feature attractor circuit (and which the participants in my UCSD course have experimented with for many years) has demonstrated that it reliably, robustly, and rapidly converges to the nearest matching pure symbol state that most closely matches (in a sort of analog Hamming metric) the initial state of the cortical region. This itself is remarkable; even though this simple model contains strong assumptions (such as a precisely and globally controlled excitation threshold being supplied to all neurons).

The latest model of the feature attractor circuit (to be published soon) eliminates these biologically unrealistic assumptions and yet retains all of the performance seen in the earlier model.

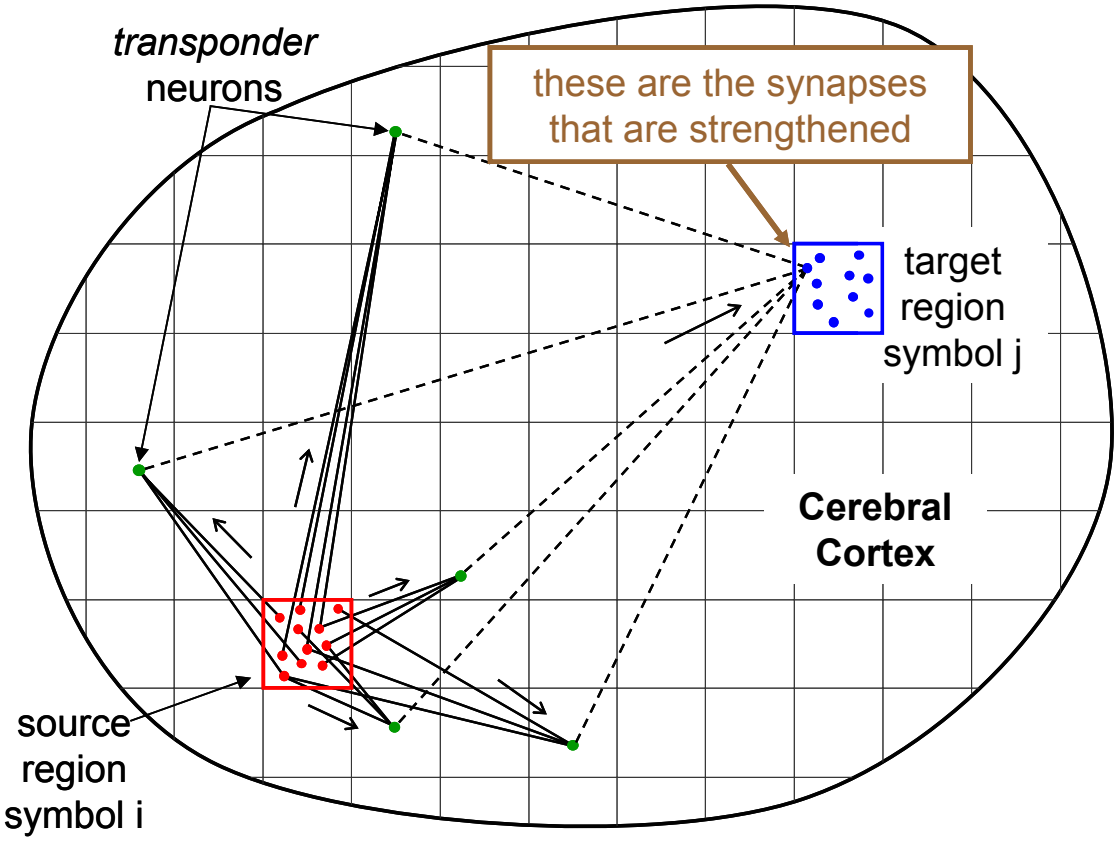

Figure 5 Cortical neuronal implementation of one item of knowledge, unidirectionally linking symbol i of the link's *source region* lexicon to symbol j of the link's *target region* lexicon. As discussed earlier, each symbol is represented by hundreds of feature detector neurons on the cortical region of its feature attractor lexicon. The hundreds of active neurons of symbol i (red dots) send their active outputs randomly (cortex is assumed to be permanently wired in a locally random pattern) to millions of excitatory pyramidal neurons. A few thousand of these millions receive a sufficient number of these inputs to become immediately and vigorously excited. These are termed *transponder neurons*. All of the (potentially tens of thousands or more) links that emanate from symbol i to symbols of all other lexicons are implemented by synapses from this one set of transponder neurons. In the instance illustrated here, we consider the small subset of transponder neurons possessing axon collaterals which just happen to synapse with one of the feature detector neurons which represent symbol j on the target region (green dots). Only a small fraction of the neurons representing symbol j (shown as blue dots), perhaps 10%, receive a sufficient number of these synapses (only connections from transponder neurons to one of these target symbol neurons are shown here). The postulated *cortical learning process* is the strengthening of these synapses to a level that, on average, equals A $\log(p(i|j)/p_0)$ + C; where A and C are fixed positive constants for all of cortex, p(i|j) is the antecedent support probability weight of this link, and p_0 is the smallest meaningful p(i|j) value. In the event that both symbols i and j are active when this link's knowledge base has been activated, these final synapses are temporarily strengthened for a period of minutes or hours. If such a strengthened link is used again during that period of wakefulness, this further enhances and prolongs this strengthening. This is the phenomenon of *short term memory*. Some short term memories that persist until the next period of sleep are then rehearsed many times – which moves them forward in the process of establishing the synaptic strengthening as permanent. Those memories which are sufficiently rehearsed on multiple, closely-spaced nights become permanently strengthened. These are the knowledge links of cerebral cortex; of which some humans may have billions.

As the formation of new triples wanes, the feature detector circuit is frozen; only to be re-animated for modification in the event of the persistent appearance of inputs that are radically different from any in the existing ensemble of triples (in which case new triples can be added – thereby expanding the set of triples without causing any harm to the already existing set of triples). In the event of severe trauma (e.g., stroke), the feature attractor is presented with nothing but novel, unfamiliar, inputs; in which case many new triples are learned to incorporate these new inputs. However, in the process, the entire capacity of the circuit is used up and all the old triples are lost. This is the cortical lexicon rebuilding process that stroke victims probably must go through in the months following the attack.

At the end of the network organization process the set of triples of the feature attractor circuit provide a set of exemplars for representing individual inputs from the information environment being supplied to the circuit as input. This set is the feature attractor circuit's *lexicon*, of which each triple in the set represents one *symbol*. Each feature detector neuron ensemble of the lexicon has a few hundred neurons that represent one symbol. The corresponding Layer VI neuron ensemble (excited by that particular feature detector ensemble) has a few tens of neurons which also represent that same symbol. The corresponding thalamic volume glomeruli that are highly excited by this Layer VI neuron ensemble also represent this same symbol. These three ensembles make up one triple; representing one symbol of the lexicon. These glomeruli then send their output back up to the original feature detector neurons representing the symbol (and/or their precedence principle predecessors), completing a stable loop. As shown in Figure 4, there is another component to the feature attractor circuit: the inhibitory reticular nucleus of the thalamus (NRT). This plays a crucial role in the operating dynamics of the feature attractor circuit; but detailed description of its function is beyond the scope of this sketch.

Further, again because of the randomness of the wiring, the individual ensembles of neurons which become highly excited together will tend to have very low pairwise overlap; even for inputs that are fairly 'similar' to one another. Except for the recent interest in sparse representations in multiple basis sets shown by information theorists, the properties of ensembles of VQ codebook vectors that are allowed to jointly become 'excited' together (because they are all fairly close to the input vector) have not been studied much. Normal VQ systems only allow the single closest codebook vector to respond to an input. However, for the past 10 years the participants in my UCSD graduate course ECE-270 (**Neurocomputing**) have been assigned a significant number of theoretical and computer simulation exercises on this subject. While these results have not yet been published (this report will now facilitate that), many of the properties described above are easy to establish.

The net result of the feature attractor circuit development process is the creation of a lexicon of a few thousand stable states representing a rich and accurate set of descriptors for the input information environment that that particular circuit is exposed to. Each time the circuit is not 'busy' (i.e., there have been no inputs for a while, or it has processed an input and descended into, and held, one of its stable states for a few tens of milliseconds and has now become 'fatigued' with that activity), any new input causes the circuit to 'describe' it by falling into its close matching symbolic state (if any). In this way, the feature attractor functions as an 'analog to symbolic converter.' During sleep, the general arousal input to cortex is suspended and feature attractors stop functioning (unless they are specifically individually selected to participate in learning during sleep). However, at all times during wakefulness, feature attractors will briefly fall into their close matching (in terms of feature attractor neuron Hamming distance) symbolic state whenever they are presented with a new input that they recognize (multiple feature attractors can be used to implement large lexicons).

Besides their function as 'analog to symbolic converters,' feature attractor neuronal circuits can carry out other duties as well. For example, when a symbol in the lexicon of a particular feature attractor is to be used as an assumed fact in a confabulation operation; it is important that the ensemble of feature detector neurons of that circuit representing the assumed fact symbol be held in an active state until the relevant knowledge base has operated (this is discussed below) and delivered the excitation from this assumed fact symbol to the symbols of the answer region(s) to which it is linked. Because of the relative delays of the links (axonal propagation speeds range from 1 to 100 m/s), some assumed fact feature attractors may have to hold their symbols active for quite a while (many tens of milliseconds).

As discussed above, feature attractors that are running free will frequently, and autonomously, switch their symbolic state as new inputs arrive. This is not acceptable when such a circuit is representing an assumed fact being used in a confabulation operation. This is where the *operation command input* of each feature attractor circuit comes into play. This operation command (the triggering of which will be discussed later) simply causes a feature attractor to hold its present symbol active as long as the command is sustained (up to perhaps a few hundred ms). This is used in setting up and momentarily locking in the assumed fact symbol inputs needed by a confabulation operation. The issuance of these feature attractor operation commands is part of the *action command sequence* that defines the particular confabulation thought process being carried out (these are the thought process analogs of a muscle contraction command sequence involved in a movement process).

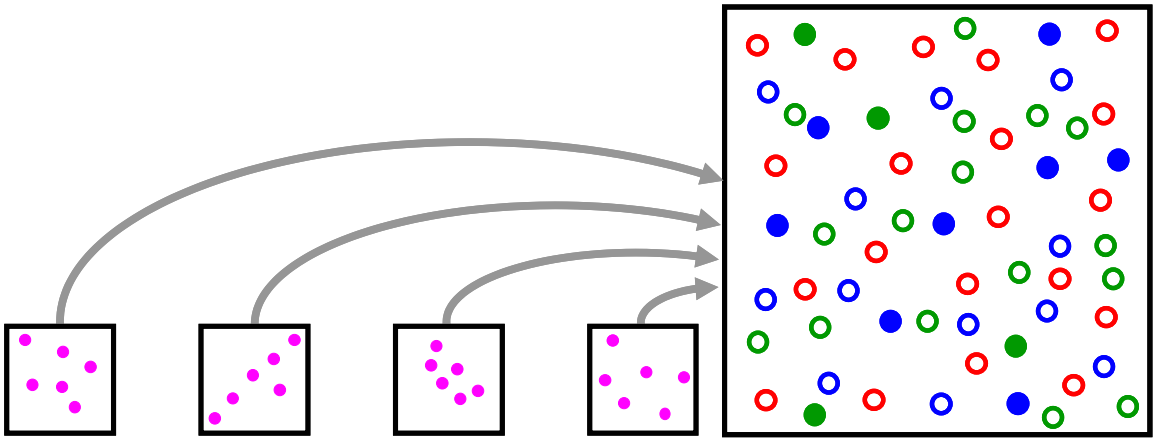

Figure 6 Cortical implementation of confabulation. Here, assumed fact symbols (each represented by hundreds of neurons, of which six are shown in pink) on four cortical source regions are delivering knowledge item link input excitation to neurons representing lexicon symbols on a cortical answer region (which has been erased and is being held static) about to carry out confabulation. For clarity, the answer region is enlarged. The answer region has thousands of symbols, each represented by a fixed set of hundreds of feature detector neurons; but here, only 20 of these feature detector neurons are shown for each of three of these symbols. One symbol's neurons are shown in red with all open circle neurons because none of these are receiving link input excitation. Green neurons representing a second symbol are receiving two links; so roughly 20% of them are being excited (the excited neurons are illustrated by filled circles). A third symbol's neurons (blue) are receiving excitation from four links and so roughly 40% of its neurons are being excited. These simultaneous knowledge base inputs require only a few tens of milliseconds to have their excitation effect. Then the answer region feature attractor operation command is suspended (see text) and this circuit snaps into the single symbol state which received the highest total excitation (the sum of the final link synapse efficacies); chosen from among those symbols that received the largest number of knowledge links (in this case four). As described in the text, after the feature attractor converges (which takes only a few tens of milliseconds), the only neurons active on the answer region will be those representing that answer lexicon symbol with the highest estimated quality. Thus, the feature attractor circuit can implement confabulation. By only partially relaxing the feature attractor control input, the highly excited symbols can have their entire complement of representing neurons excited. But the competition process that causes elimination of all but the top answer is suppressed in this case. This causes the symbols of the expectation to be left fully excited for a brief moment. Additional link input(s) are then immediately brought in and the control input is dropped; causing the feature attractor to snap into the best matching symbol from among the expectation symbols. This is how expectations are formed and used.

Its operation command can also be used to cause a feature attractor circuit to carry out confabulation (i.e., to carry out the function of an answer lexicon). The first step in this process is to *erase* or *clear* the circuit of any symbol that it may be expressing or may be in the process of activating. This is done by sending the operation command twice in succession; with the second command input sustained. The first input freezes the circuit's state. The immediate interruption and resumption of the command then causes the circuit to assume a *passive state* in which the attraction dynamics of the circuit are suspended for the duration of the second operation command signal. In this passive state, feature detector neurons of the cortical region of the circuit can receive

input and become excited; but the rest of the feature attractor circuit (in particular, the thalamic volume of the circuit) will not respond to these excitations. It is during this passive state (which usually only exists for a tiny fraction of a second and, at most, a second or two) that the inputs from the assumed fact symbols are received by the feature detector neurons. As soon as the feature detector neurons have received (and processed through their precedence hierarchy) all of these assumed fact excitations, the operate command is momentarily either fully suspended (if the best answer is desired) or 'lightened up' (if an expectation is desired). The resulting best answer symbol, or an expectation set of symbols, can then be used as an input to the next stage of mental information processing (how confabulation is implemented by a feature attractor is described below in subsection 3.3). The theory holds that this confabulation operation is the only information processing that goes on in the cerebral cortex. All mental information processing is constructed out of combinations of such processes. As will be described below, each time an answer is determined (but not an expectation), a new action command is issued. Thus, cortical processing can implement *branching operations*; in the sense that the selection of subsequent thought commands (confabulation process commands) can depend on the outcome of a previous confabulation. It is this property that makes cortex a *universal computer* in roughly the Turing sense.

As mentioned, the theory predicts that each cognitive processing step (confabulation operation) will be slightly preceded by a distinctive 'double pulse' command input to clear the feature attractor. This is perhaps the widely reported and seemingly ubiquitous 'cognition-related field potential oscillation' that seems to slightly precede cortical cognitive processing [Konig; Stryker].

3.2 Knowledge Item Implementation in Cortex

Figure 5 illustrates the manner in which the theory envisions an individual item of knowledge being implemented in cortex. These links utilize a two-step transmission process through the unchanging, and locally randomly wired, cortical axonal network. My preliminary mathematical and computer simulation studies of a simplified model of this design (and related models built by participants in my UCSD ECE-270 course), which have not yet been published, indicates that this scheme will indeed work: an individual symbol can implement tens of thousands of links to other symbols on other regions without interference or link failure. Having many symbols (each on a different region) transmitting at the same time to the same target region (where confabulation will be carried out) does not seem to cause interference problems either. Another characteristic of this design is its robust tolerance of gradual random symbol-representation feature detector neuron and transponder neuron death.

A key insight from these studies of knowledge item implementation is that only a small fraction (perhaps 10%, although this probably varies from one functional area of cortex to another) of each target symbol's neurons will receive knowledge link inputs from the source symbol's transponder neurons. At first, this inescapable conclusion seemed very bad for the theory. But it is actually very good! This is how the arithmetic of feature attractor circuit operation is kept precise (which it *must* be): by replacing dependence on individual neuron summation processes with a process that simply counts up the total average excitation being delivered to different sized groups of neurons (smaller groups have no chance in the confabulation competition and the largest groups compete on total distributed excitation). This is now explained.

3.3 How Feature Attractors Carry Out Confabulation

Figure 6 illustrates how confabulation is hypothesized to be implemented by an answer feature attractor circuit. Each symbol of the answer circuit lexicon is receiving an integer number of links (0, 1, 2, 3, 4, etc.). In the example of Figure 6, some of the answer circuit symbols (such as the symbol with blue neurons illustrated) are receiving links from all four of the assumed fact symbols shown. Assuming that the feature attractor circuit is being used to obtain the best answer (as opposed to forming and using an expectation, a mode of operation which is discussed separately below), those answer circuit symbols that are receiving zero, one, two, or three links have no chance whatsoever of being the symbol to which the circuit will converge when the operate command is released; no matter what their total excitations. This is because the number of representing neurons of these symbols that are receiving excitation (roughly 0%, 10%, 20%, or 30%) are simply not going to successfully compete against any symbol that has 40% of its neurons being excited, as their signals propagate through the cortex-to-thalamus-to-cortex loop of the circuit. Thus, the competition for the winning symbol will be exclusively among those which received the largest number of links (in this case, four).

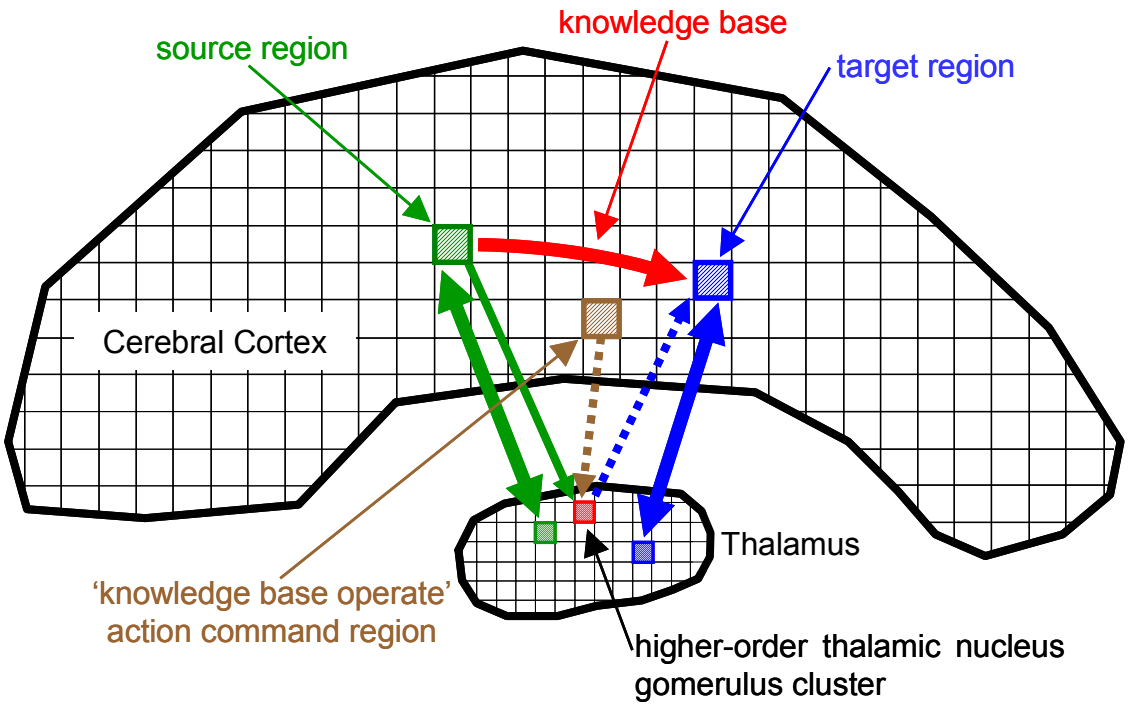

Figure 7 Speculation on how individual knowledge base control works. Here, an answer (i.e., target) region – shown in blue – is to receive excitation from knowledge links emanating from an active symbol on a source region – shown in green. In order for the excitation from these links to have any effect on the symbol neurons they synapse with in the target region, the target region must receive a *preparatory* signal in close temporal proximity to the knowledge link inputs. This preparatory signal – shown as a narrow dotted blue arrow -- originates from a cluster of glomeruli (shown in red in the thalamus) in a *higher-order* thalamic nucleus (see text). The preparatory signal is nothing but a relay (by the cluster) of a layer VI signal sent from the source region whenever any one of its symbols has become active immediately following a confabulation operation carried out on that region. However, the thalamic cluster will only relay this signal if it has been 'switched on' (modulated) by a *thought command* signal (shown as a dashed brown line) which originates as a layer V output signal in yet a third cortical region. If this thought command signal has been issued by the brown region, then the red cluster's glomeruli are modulated on and the (driving) input (narrow green arrow) from the green region is immediately relayed to the blue region (dotted blue line). If this thought command signal has not been issued by the brown region, then the red cluster's glomeruli are not modulated on and the green signal is not relayed – in which case the blue region does not respond to the incoming knowledge link excitation from the green region. This design requires one unique thalamic cluster for each of the hundreds of thousands of knowledge bases in human cortex – which is exactly what the theory postulates. Further, it requires that each paired knowledge base and cluster have one or more control regions (there is no reason there cannot be more than one for each knowledge base) assigned to it. These regions provide the thought control action command outputs that, by the above mechanism, enable the operation of the links of that knowledge base. Without this individual control of each knowledge base, pairs of regions that are linked by a knowledge base could never function at the same time without invoking this knowledge. This restriction would cause frequent interference between thought processes.

CRITICAL MATERIAL

The *total excitation* of each contending target symbol is defined to be the sum of the average efficacies of its link implementation synapses. In this instance, for contending answer symbol ε_5, this will be (ignoring random overlaps between the excited target neuron subsets) a sum of the form $[A \log(p(\alpha_1|\varepsilon_5)/p_0) + C] + [A \log(p(\beta_2\varepsilon_5)/p_0) + C] + [A \log(p(\gamma_3|\varepsilon_5)/p_0) + C] + [A \log(p(\delta_4|\varepsilon_5)/p_0) + C]$ (where A, p0, and C are explained in the caption of Figure 5, and α_1, β_2, γ_3, and δ_4 are the symbol indices of the assumed fact symbols on source regions 1, 2, 3, and 4 – as shown in Figure 6). This sum can be mathematically re-expressed as $A \log[p(\alpha_1|\varepsilon_5) \cdot p(\beta_2\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)] - 4 \log(p_0) + 4\ C$. However, since the logarithm is a strictly monotonically increasing function and the constant A is positive, the ordering of these total excitations among the contending

symbols will be exactly the same as the ordering of their $p(\alpha_1|\varepsilon_5) \cdot p(\beta_2\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)$ values (there estimated qualities). Thus, the answer feature attractor will carry out confabulation; ending (after the circuit dynamically snaps into the most excited pure state) with all the neurons representing the symbol ε_5 with the highest estimated quality in an active state (and all other feature detector neurons inactive) on its cortical region at the end of its operation period following release of the circuit's operation command.

Ignoring the propagation transit times of the knowledge link signals (which, for some long links, might need to leave their source regions 10 to 30 ms before confabulation begins), an answer feature attractor circuit can probably carry out a confabulation operation in well under 100 ms. Consider that one such operation might well involve five or six, or sometimes more, assumed facts: each sending tens, to as many as thousands, of knowledge links to hundreds or even thousands of answer region symbols. That the neurons of a feature attractor circuit (which can each be simplistically viewed as 'processors with 1 kHz clocks') can implement the effective equivalent of the millions of calculations needed to determine the best answer symbol in a tiny fraction of a second is quite impressive. This is the method by which the brain can carry out cognition in time intervals only slightly longer than those required for the propagation of the involved axonal signals. [That brains seem to be able to do this is a widely noted fact – consider that the player of a simulated 'gunfighter' game can 'shoot' a bad person (who suddenly appears in the 'wild west' street video scene being observed by the player) with a total delay of only a few hundred milliseconds – not much longer than the axonal transport delay time from eye to sensory brain to motor brain to arm and trigger finger muscles.]

When the answer feature attractor control input is not completely released at the conclusion of knowledge link input, but only relaxed, the feature attractor completes one round trip of internal interaction, but is inhibited from selecting a single winning token. All of the neurons representing symbols receiving the maximum number of links become excited and all other neurons are suppressed. This is the neural implementation of an expectation. Additional assumed fact links can then be brought in (over the next few tens of milliseconds) and that symbol belonging to the expectation which receives the greatest total additional excitation input will (when the control input is then completely released) have all its neurons become active on the answer region. The neurons of all other symbols will effectively be silent. If the supplemental input does not target any of the expectation symbols the circuit collapses into a null state (meaning that no symbol of the expectation was confirmed).

In this way, an expectation can be used to confine or restrict a confabulation operation to a pre-selected set of *candidate answers* – namely, those belonging to the expectation. For many applications (for example, instant and automatic segmentation of a single, complicated, attended – 'expected' – object from an auditory or visual scene), this is the essential operation. Expectation can be used to explain the long-known, but paradoxical, observation that often, at any given moment, higher sensory areas only seem to be responding to inputs from a single attended object. The theory explains this as the use of expectation to 'clip out' the attended object from the other elements of the "blooming, buzzing, world" (to quote Professor George Miller of Princeton) at the lowest levels of cortical sensory processing; thereby allowing the subsequent portions of cortex to concentrate on this isolated object.

3.4 Knowledge Base Control

As just mentioned, each knowledge base (group of all knowledge links from one particular cortical lexicon to a second) must be deliberately activated for it to function properly. In essence, this means that a *command signal* must be sent for a knowledge base (of which mature human cortex might have hundreds of thousands) in order for its links to be able to function correctly. I don't know exactly how this mechanism works; but suspect that the process involves what Sherman and Guillery [Sherman and Guillery] term *higher-order* thalamic nuclei (which make up a large fraction of thalamus and do not contain any of the thalamic volumes involved in feature attractor circuits). The idea (see Figure 7) is that a special signal sent from layer V of a cortical region (see below) to a particular small cluster of glomeruli (in a higher-order nucleus) uniquely dedicated to that knowledge base *enables* the knowledge links (in one direction) between two particular cortical regions to operate. If no such signal is sent to this cluster, then the links will not operate. The speculation of the theory is that such a special signal from a layer V of a cortical region to the cluster might act to 'facilitatively modulate'

this cluster. These might be some of the *modulating inputs* from cortex that [Sherman and Guillery] speak of. Then, when the source region of a knowledge base launches its link signals a short time later, a subset of the corresponding active layer VI neuron ensembles of this source region send *driving* outputs to this same cluster – which, because it has recently been facilitated, vigorously sends these signals on to the target region. Presumably, these driving outputs to the target cortical region (which will arrive very close to the time of the actual link excitations, due to the length similarities of the involved axons) go to layer IV and somehow 'prepare' or 'command' the feature detector neurons there to process the corresponding link excitation inputs. Without this *preparatory* signal, the target region will presumably not respond to the knowledge link excitations from this particular source region.

The control of individual knowledge bases is important because two regions which have direct knowledge links may be involved in a thought process at the same time, but not with each other. It would be disastrous if all links (which go from each symbol to symbols on perhaps hundreds of different lexicons) were always 'functional.' Obviously, this scheme would require that there be a unique dedicated cluster of "higher order" thalamic glomeruli for each knowledge base. This cluster would have to receive its modulation input from whatever cortical region that controls it and would have to receive its driving input from the source region of that knowledge base. The output of the cluster's glomeruli would go to the target region of the knowledge base; where it acts to facilitate reception of link excitation. One suspects that the 'self-timing' of the signals in this design is critical: the preparatory input arriving perhaps exactly at, or slightly before, the link excitation. This may explain why some cortico-cortical axons become myelinated: to facilitate the precise relative timing of the links and their preparatory signals. As is the case with the feature attractor circuit design, these speculations about knowledge base control generally seem to fit the (quite limited) known facts about the wiring of cortex and thalamus.

3.5 Action Command Generation: Every Moment From Wake-Up to Sleep

The brown elements of Figure 4 (not mentioned in its caption) and the knowledge base control commands of Figure 7 illustrate the theory's explanation for how cerebral cortex controls movement and cognition. The basic idea is that each time a symbol is made active by a confabulation operation (either an immediate best answer or the symbol that results from successful resolution of an expectation) on an answer feature attractor circuit, a small collection of neurons in layer V of that region, which are permanently (but perhaps not uniquely) paired with that symbol, also automatically and instantly become active also. These layer V neurons generally send their output axons to subcortical structures (termed *action nuclei* by the theory) located in many places throughout the central nervous system. These layer V outputs are termed *action commands*. Typically, tens to hundreds of such action commands are generated each second we are awake.

In principle, each lexicon symbol can have its own unique paired action command. But in many cases, multiple symbols trigger the same action command. The special 'brown' signals of Figure 7 that control knowledge bases are action commands.

Many of the action commands issued by feature attractor circuits located in primary motor cortex (and in some other brain areas, e.g., perhaps the frontal eye fields) proceed directly to *motor nuclei* of the brainstem and spinal cord. Thus, motor nuclei are a type of action nucleus. Namely, they are responsible for the control of *movement processes* – one type of action. Action commands from feature attractors not located in primary motor cortex typically proceed to subcortical *thought nuclei*; which (the theory postulates) are responsible for controlling cortical feature attractors and knowledge bases. Many of these thought nuclei are clusters of glomeruli in the higher order nuclei of the thalamus, as shown in Figure 7. But others are located in the brainstem and elsewhere.

That a vast majority of cortex would be involved in issuing thought action commands, as opposed to movement action commands, makes sense because there are many more feature attractors and knowledge bases than muscles. So it is no wonder that a much larger portion of cortex is devoted to producing such *thought process control* action commands.

Most action commands represent 'low-level housekeeping functions' that are executed reflexively whenever a particular symbol (often one of a large set of symbols that will elicit the same action command) becomes active on the region. For example, whenever a circuit has been used as the answer region of a confabulation process that is recalling a stored action sequence (these are typically located in frontal cortex ahead of the central sulcus), that circuit (and others which always function with it) must then be immediately erased and prepared for generating the next sequence symbol. This is an action command that is issued along with the expression of the current action sequence symbol. Overriding such reflexive thought progressions is possible; but generally involves shutting off all cortical arousal in a general area via action commands issued to brainstem thought nuclei. The result is a momentary freezing of function as a new thought process stream is inaugurated.

Many action commands (particularly those associated with selection of high-level behaviors) cannot be generated immediately; but must first be vetted by the basal ganglia. There is much more to say about action command generation (little of which I know – e.g., you want to know where the feature attractor operation commands come from and how they function; well so do I); but the above gives an idea of the basic view of the theory).

4. Discussion

Feature attractor networks and the neurons which implement items of knowledge begin dying, slowly and randomly, as soon as they are formed. A compelling aspect of this theory is that both feature attractor circuits and knowledge links are quite insensitive to this random loss of tissue. Computer experiments with feature attractor models suggest that half of the initial complement of neurons can randomly die without effecting overall correct functioning. Numerical studies of knowledge links suggest a similar, although perhaps somewhat lower percentage, failure tolerance. However, I suspect that there is quite a bit more to this story. In particular, the lower failure tolerance of links probably means that long-disused knowledge items will suffer eventual loss because of the gradual erosion of the transponder neurons and target symbol neurons. Before complete loss, sometimes the use of multiple knowledge links aimed at the same symbol (e.g., when we use mnemonics to recall an almost-forgotten item of knowledge) can sometimes 'assist' a nearly-lost link in recalling its target token. However, if a knowledge item is occasionally used, neurons which were previously not as well connected (or which have been newly created and randomly wired up), can be recruited and added to the circuit to restore the link to full functionality. This is why we forget knowledge after very long disuse, but retain knowledge that is at least occasionally used.

The neuron ensemble triple associated with a symbol in a feature attractor circuit probably also slowly degrades, if unused (not just through neuron death, but also because of the circuit's natural propensity to adjust to its long-term statistical environment). Thus, over long periods of disuse (typically many years) a symbol can disappear from the circuit's lexicon. In my own case, total lack of contact with ameretto cookies (which I loved in early childhood) from about age 7 to age 28 almost extinguished that odor representation symbol from one of my smell lexicons. Upon experiencing the odor again after more than two decades, it took considerable effort at first to 'recall' it and place it in context. After many seconds of concentrated effort (with the effort being led by the limbic system – the first realization being that this was a 'very happy' odor); everything came rushing back. Subsequently, neither this symbol, nor its knowledge links, have ever again been allowed to wilt.

A conclusion of this theory is that although perhaps hundreds of them are used for storing each individual item of knowledge (token to token link), and although many synapses probably go unused, the 10^{14} to 10^{15} synapses of the mature human cerebral cortex [Mountcastle; Nichols; Steward] probably provide us with a storage capacity measured in billions of items of knowledge. This implies that children (and adults too!) probably accumulate millions of new items of antecedent support knowledge every day. Sorting each day's short and medium term memories and converting selected ones into a more permanent form is a huge job. It is no wonder that we must sleep a third of the time.

To appreciate the vast storage capacity of your cerebral cortex; imagine for a moment that you are being asked a long series of detailed questions about the kitchen in your home. Describe all the spoons and where they are

kept; then the forks, then the drinking glasses, and so on. Describe how you select and employ each item, where and when you obtained it, and some memorable occasions when it was used. Obviously, such a process could go on for tens of hours and still turn up lots of new information. Now consider that you could probably answer such detailed questions for thousands of mental arenas.

Given sufficient time and effort, humans and some other vertebrates can answer certain novel questions for which there are no pre-existing stored thought processes. A raven (presumably a confabulator of some sort) can fashion a hook at one end of an initially straight length of wire and then use it to pull a small basket of food up and out of a long clear plastic tube [Weir]. Humans have proven the Four Color Theorem and Fermat's Last Theorem, and developed a theory of quantum electrodynamics. But these feats (excellent *simulations* of reasoning by labored use of confabulation within highly constrained, and natively 'logical,' cognitive domains) are highly misleading if used as hints about the workings of ordinary thought. [When you park your car you do not reason – you simply confabulate your way into the available spot.] It was perhaps such misleading hints that caused this theory not to have been discovered many years ago.

This completes this introductory sketch of the cortical theory. Obviously, there are many missing pieces and details. But the general picture of a unified mechanism for knowledge acquisition and use seems compelling and correct. It has exactly all the right qualitative characteristics and it answers the most difficult questions (such as how cognition can be carried out so fast and reliably over time using hardware that is slow and constantly dying). Further, the theory offers a clear picture of how it could be tested. Either there are feature attractor networks and knowledge links or there aren't.

Finally, notwithstanding all the arguments presented here in favor of confabulation being the underlying mechanism of cognition, many will not accept this hypothesis without more evidence. For example, more computer thinking experiments. Well, these exist and they are strongly supportive of the theory's claims. These results are beginning to be published separately. Overall, the main purpose of this report is to provide general background for more focused reports to follow.

Acknowledgements

Thanks to Adrian T. Fan, Katherine G. Mark, Robert W. Means, and Syrus C. Nemat-Nasser for implementing the computer thinking experiments of the second section and to Fair Isaac and ONR for long-term research support. Domestic cat Zeus Hecht-Nielsen provided key observational insights that led to this theory.

References (including some work that influenced the development of this theory)

Abeles, M. (1991) Corticonics. Cambridge, UK: Cambridge Univ. Press.

Amari, S. (1989) Characteristics of sparsely encoded associative memory. Neural Networks 2: 451–457.

Amari, S. (1974) A method of statistical neurodynamics. Biological Cybernetics 14: 201–215.

Amari, S. (1972) Learning patterns and pattern sequences by self-organizing nets of threshold elements. IEEE Trans. Comput. C-21: 1197–1206.

Amit, D. (1989) Modeling Brain Function: The World of Attractor Networks. Cambridge, UK: Cambridge Univ. Press.

Anderson, J.A. (1968) A memory storage model utilizing spatial correlation functions. Kybernetik 5: 113–119.

Anderson, J.A. (1972) A simple neural network generating an interactive memory. Mathematical Biosciences 14: 197–220.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., Jones, R.S. (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. Psych. Rev. 84: 413–451.

Bain, A. (1873) *Mind and Body: The Theories of Their Relation*. London: Henry King.

Borisyuk, R., Borisyuk, G., Kazanovich, Y. (1998) Synchronization of neural activity and information processing. Behavioral and Brain Sciences 21: 833–844.

Braitenberg V., Schuz, A. (1998) Cortex: Statistics and Geometry of Neuronal Connectivity, Second Edition. Berlin: Springer-Verlag.

Carpenter, G.A., Grossberg, S. (Eds.) (1991) Pattern Recognition by Self-Organizing Neural Networks. Cambridge, MA: MIT Press.

Carpenter, G.A., Grossberg, S. (1988) Adaptive Resonance. In: Grossberg, S. (Ed.) Neural Networks and Natural Intelligence. Cambridge, MA: MIT Press, 251-315.

Cohen, M.A., Grossberg, S. (1983) Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. IEEE Trans. Sys. Man and Cyber. 13: 815–826.

Cowan, W.M., Sudhof, T.C., Stevens, C.F. (Eds.) (2001) Synapses. Baltimore, MD: Johns Hopkins Univ. Press.

Creutzfeldt, O.D. (1995) *Cortex Cerebri*. Oxford University Press.

Crick, F.H.C. (1984) Function of the thalamic reticular complex: The searchlight hypothesis. Proc. Nat. Acad. Sci. 81: 4586–4590.

Crossman, A. R. and Neary, D. (1995) *Neuroanatomy*. New York: Pearson Professional Ltd.

Desai, N. S., Cudmore, R. H., Nelson, S. B., and Turrigiano, G. G. (2002) Critical periods for experience-dependent synaptic scaling in visual cortex. Nature Neurosci. 8: 783-789.

Fain, G.L. (1999) Molecular and Cellular Physiology of Neurons. Cambridge, MA: Harvard Univ. Press.

Fries, P., Reynolds, J.H., Rorie, A.E., Desimone, R. (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291: 1560–1563.

Fukushima, K., Miyake, S. Ito, T. (1983) Neocognitron: a neural network model for a mechanism of visual pattern recognition, IEEE Transactions on Systems, Man, and Cybernetics SMC-13: 826-834

Fukushima, K., Miyake, S. (1978) A self-organizing neural network with a function of associative memory: Feedback-type Cognitron. Biological Cybernetics 28: 201–208.

Fukushima, K. (1975) Cognitron: A self-organizing multilayered neural network. Biological Cybernetics 20: 121–136.

Grossberg, S., Williamson, J.R. (2001) A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual groupings and learning. Cerebral Cortex 11: 37–58.

Grossberg, S. (1999) How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex. Spatial Vision 12: 163–186.

Grossberg, S., Mingolla, E., Ross, W.D. (1997) Visual brain and visual perception: How does the cortex do perceptual grouping? Trends in Neurosciences 20: 106– 111.

Grossberg, S. (1997) Cortical dynamics of three-dimensional figure-ground perception of two-dimensional patterns. Psych. Rev. 104: 618–658.

Grossberg, S. (1995) The attentive brain. American Scientist 83: 438–449.

Grossberg, S. (Ed.) (1987) The Adaptive Brain, Volumes I and II. Amsterdam: Elsevier.

Grossberg, S. (Ed.) (1982) Studies of Mind and Brain. Norwell, MA: Kluwer.

Grossberg, S. (1976) Adaptive pattern classification and universal recoding. Biological Cybernetics 23: 121–134.

Grossberg, S. (1968) Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. Proceedings of the National Academy of Sciences 59: 368–372.

Haines, K., Hecht-Nielsen, R. (1988) A BAM with increase information storage capacity. Proceedings, 1988 International Conf. on Neural Networks, Piscataway NJ: IEEE Press, I-181–I-190.

Hall, Z.W. (1992) Molecular Neurobiology. Sunderland, MA: Sinauer.

Hebb, D. (1949) The Organization of Behavior. New York: Wiley.

Hecht-Nielsen, R., McKenna, T. (Eds.) (2003) Computational Models for Neuroscience. (Springer-Verlag, London, 2003), pp. 85–124.

Herculano-Houzel, S., Munk, M.H.J., Neuenschwander, S., Singer, W. (1999) Precisely synchronized oscillatory firing patterns require electroencephalographic activation. J. Neurosci. 19: 3992–4010.

Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. 79: 2554–2558.

James, W. (1890) *The Principles of Psychology*. New York: Henry Holt & Co.

Jones, E.G. (1985) The Thalamus. New York: Plenum Press.

Kainen, P.C., Kůrková, V. (1993) On quasiorthogonal dimension of euclidean space. Applied Math Lett. 6: 7–10.

Kohonen, T. (1995) Self-Organizing Maps. Berlin: Springer-Verlag.

Kohonen, T. (1984) Self-Organization and Associative Memory. Berlin: Springer-Verlag.

Kohonen, T. (1972) Correlation matrix memories. IEEE Transactions on Computers C21: 353–359.

Konig, P., Engel, A.K., Roelfsema, P.R., Singer, W. (1995) How precise is neuronal synchronization. Neural Computation 7: 469–485.

Kosko, B. (1988) Bidirectional associative memories. IEEE Trans. On Systems, Man, and Cybernetics SMC-18, 49-60.

Kryukov, V.I., Borisyuk, G.N., Borisyuk, R.M., Kirillov, A.B., Kovalenko, E.I. (1990) Metastable and unstable states in the brain. In: R.L. Dobrushin, V.I. Kryukov, A.L. Toom (Eds.) Stochastic Cellular Systems. Manchester, UK: Manchester Univ. Press.

Ledoux, M. (2001) The Concentration of Measure Phenomenon. Providence, RI: American Mathematical Society.

Marder, E. and Prinz, A.A. (2003) Current compensation in neuronal homeostasis. Neuron 37: 2-4.

Marder, E. and Prinz, A.A. (2002) Modeling stability in neuron and network function: the role of activity in homeostasis. BioEssays 24: 1145-1154.

Miller, G. A. (1996) The Science of Words. New York: Scientific American Library.

Mountcastle, V.B. (1998) Perceptual Neuroscience: The Cerebral Cortex. Cambridge, MA: Harvard Univ. Press.

Nakano, K. (1972) Associatron - a model of associative memory. IEEE Transactions on Systems, Man, and Cybernetics SMC-2: 380–388.

Nicholls, J.G., Martin, A.R., Wallace, B.G., Fuchs, P.A. (2001) From Neuron to Brain, Fourth Edition. Sunderland, MA: Sinauer.

Nilsson, N.J. (1998) Artificial Intelligence: A New Synthesis. San Francisco: Morgan Freeman Publishers.

Nilsson, N.J. (1965) Learning Machines. New York: McGraw-Hill.

Nolte, J. (1999) The Human Brain, Fourth Edition. St. Louis, MO: Mosby.

Oja, E. (1980) On the convergence of an associative learning algorithm in the presence of noise. International Journal of Systems Science 11: 629–640.

Palm, G. (1980) On associative memory. Biol. Cybern. 36: 19–31.

Pearl, J. Causality. (2000) Cambridge University Press.

Sagi, B., Nemat-Nasser, S.C., Kerr, R., Hayek, R., Downing, C., Hecht-Nielsen, R. (2001) A biologically motivated solution to the cocktail party problem. Neural Computation 13: 1575-1602.

Sejnowski, T.J., Destexhe, A. (2000) Why do we sleep? Brain Research 886: 208– 223.

Sherman, S.M., Guillery, R.W. (2001) Exploring the Thalamus. San Diego, CA: Academic Press.

Sommer, F.T., Palm, G. (1999) Improved bidirectional retrieval of sparse patterns stored by Hebbian learning. Neural Networks 12: 281–297.

Steinbuch, K. (1965) Automat und Mensch, Third Edition. Heidelberg: Springer-Verlag.

Steinbuch, K. (1963) Automat und Mensch, Second Edition. Heidelberg: Springer-Verlag.

Steinbuch, K. (1961a) Automat und Mensch. Heidelberg: Springer-Verlag.

Steinbuch, K. (1961b) Die lernmatrix. Kybernetik 1: 36–45.

Steinbuch, K., Widrow, B. (1965) A critical comparison of two kinds of adaptive classification networks. IEEE Transactions on Electronic Computers October: 737–740.

Steinbuch, K., Piske, U.A.W. (1963) Learning matrices and their applications. IEEE Transactions on Electronic Computers December: 846–862.

Steward, O. (2000) Functional Neuroscience. New York: Springer-Verlag.

Stryker, M.P. (2001) Drums keep pounding a rhythm in the brain. Science 291: 1506– 1507.

Thorpe, S., Delorme, A., Van Rullen, R. (2001) Spike-based strategies for rapid processing. Neural Networks 14: 715–725.

Steriade, M. and Llinás, R. R. (1988) The functional states of the thalamus and the associated neuronal interplay, Physiological Reviews 68: 649-742.

Steriade, M., Jones, E. G., and Llinás, R. R. (1990) Thalamic Oscillations and Signaling. New York: John Wiley & Sons.

Steriade, M., McCormick, D. A., and Sejnowski T. J. (1993) Thalamocortical oscillations in the sleeping and aroused brain, Science 262: 679-685.

Steriade, M. (1996) Arousal: Revisiting the reticular activating system, Science 272: 225-226.

Taylor, J. G. (1996) A competition for consciousness? Neurocomputing 11: 271-296.

Taylor, J. G. (1996) A neural network model of conscious and unconscious perception, Biological Cybernetics 75: 59-72.

Y. Z. Tsypkin, Y. Z. (1973) Foundations of the Theory of Learning Systems. New York: Academic Press.

Turrigiano, G. G., Nelson, S. B. (2004) Homeostatic plasticity in the developing nervous system. Nature Rev. Neurosci. 5: 97-107.

Turrigiano, G. G., Nelson, S. B. (2000) Hebb and homeostasis in neuronal plasticity. Curr. Opin. Neurobiol. 10: 358-364.

Turrigiano, G. G., Leslie, K. R. Desai, N. S. Rutherford, L. C. Nelson, S. B. (1998) Activity-dependent scaling of quantal amplitude in neocortical pyamidal neurons. Nature 391: 892-895.

Ukhtomsky, A.A. (1966) Dominanta (in Russian). Leningrad: USSR Academy of Sciences.

von der Malsburg, C. (1981) The correlation theory of brain function. Internal Report 81-2, Max-Planck-Inst. for Biophysical Chemistry.

Weir, A.A.S., Chappell, J., Kacelnik, A. (2002) Shaping of hooks in New Caledonian crows. Science 297: 981.

Widrow, B., Hoff, M.E. (1960) Adaptive switching circuits. 1960 IRE WESCON Convention Record, New York: Institute of Radio Engineers, 96-104.

Wilkes, A.L., Wade, N.J. (1997) Bain on neural networks. Brain and Cognition 33: 295-305.

Willshaw, D.J., Buneman, O.P., Longuet-Higgins, H.C. (1969) Non-holographic associative memory. Nature 222: 960–962.

Zadeh, L.A. (1965) Fuzzy sets. Information and Control, 8, 338-353.

Zador, P. (1963) Development and Evaluation of Procedures for Quantizing Multivariate Distributions, PhD Dissertation, Palo Alto, CA: Stanford University.

APPENDIX

Derivation of Equation Relating Ideal Quality and Estimated Quality

Using the identity p(abcde) = p(a|bcde) · p(b|cde) · p(c|de) · p(d|e) · p(e) [Pearl] and using the fact that ANDed events commute, we can write the quantity $p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$ in all four of the following ways:

$$p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5) \equiv p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\varepsilon_5)$$
$$= p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\varepsilon_5) = p(\alpha_1|\beta_2\gamma_3\delta_4\varepsilon_5) \cdot p(\beta_2|\gamma_3\delta_4\varepsilon_5) \cdot p(\gamma_3|\delta_4\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)$$

$$p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$$
$$= p(\beta_2\gamma_3\delta_4\alpha_1\varepsilon_5)/p(\varepsilon_5) = p(\beta_2|\gamma_3\delta_4\alpha_1\varepsilon_5) \cdot p(\gamma_3|\delta_4\alpha_1\varepsilon_5) \cdot p(\delta_4|\alpha_1\varepsilon_5) \cdot p(\alpha_1|\varepsilon_5)$$

$$p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$$
$$= p(\gamma_3\delta_4\alpha_1\beta_2\varepsilon_5)/p(\varepsilon_5) = p(\gamma_3|\delta_4\alpha_1\beta_2\varepsilon_5) \cdot p(\delta_4|\alpha_1\beta_2\varepsilon_5) \cdot p(\alpha_1|\beta_2\varepsilon_5) \cdot p(\beta_2|\varepsilon_5)$$

$$p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)$$
$$= p(\delta_4\alpha_1\beta_2\gamma_3\varepsilon_5)/p(\varepsilon_5) = p(\delta_4|\alpha_1\beta_2\gamma_3\varepsilon_5) \cdot p(\alpha_1|\beta_2\gamma_3\varepsilon_5) \cdot p(\beta_2|\gamma_3\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5)$$

Multiplying these equations together gives:

$$[p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)]^4 = [p(\alpha_1|\beta_2\gamma_3\delta_4\varepsilon_5) \cdot p(\beta_2|\gamma_3\delta_4\varepsilon_5) \cdot p(\gamma_3|\delta_4\varepsilon_5)]$$
$$\cdot [p(\beta_2|\gamma_3\delta_4\alpha_1\varepsilon_5) \cdot p(\gamma_3|\delta_4\alpha_1\varepsilon_5) \cdot p(\delta_4|\alpha_1\varepsilon_5)]$$
$$\cdot [p(\gamma_3|\delta_4\alpha_1\beta_2\varepsilon_5) \cdot p(\delta_4|\alpha_1\beta_2\varepsilon_5) \cdot p(\alpha_1|\beta_2\varepsilon_5)]$$
$$\cdot [p(\delta_4|\alpha_1\beta_2\gamma_3\varepsilon_5) \cdot p(\alpha_1|\beta_2\gamma_3\varepsilon_5) \cdot p(\beta_2|\gamma_3\varepsilon_5)]$$
$$\cdot [p(\alpha_1|\varepsilon_5) \cdot p(\beta_2|\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)].$$

Applying Bayes law to all of the quantities in the first four parentheses yields:

$$[p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)]^4 = [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\beta_2\gamma_3\delta_4\varepsilon_5) \cdot p(\beta_2\gamma_3\delta_4\varepsilon_5)/p(\gamma_3\delta_4\varepsilon_5) \cdot p(\gamma_3\delta_4\varepsilon_5)/p(\delta_4\varepsilon_5)]$$
$$\cdot [p(\beta_2\gamma_3\delta_4\alpha_1\varepsilon_5)/p(\gamma_3\delta_4\alpha_1\varepsilon_5) \cdot p(\gamma_3\delta_4\alpha_1\varepsilon_5)/p(\delta_4\alpha_1\varepsilon_5) \cdot p(\delta_4\alpha_1\varepsilon_5)/p(\alpha_1\varepsilon_5)]$$
$$\cdot [p(\gamma_3\delta_4\alpha_1\beta_2\varepsilon_5)/p(\delta_4\alpha_1\beta_2\varepsilon_5) \cdot p(\delta_4\alpha_1\beta_2\varepsilon_5)/p(\alpha_1\beta_2\varepsilon_5) \cdot p(\alpha_1\beta_2\varepsilon_5)/p(\beta_2\varepsilon_5)]$$
$$\cdot [p(\delta_4\alpha_1\beta_2\gamma_3\varepsilon_5)/p(\alpha_1\beta_2\gamma_3\varepsilon_5) \cdot p(\alpha_1\beta_2\gamma_3\varepsilon_5)/p(\beta_2\gamma_3\varepsilon_5) \cdot p(\beta_2\gamma_3\varepsilon_5)/p(\gamma_3\varepsilon_5)]$$
$$\cdot [p(\alpha_1|\varepsilon_5) \cdot p(\beta_2|\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)].$$

Assuming that all of the probabilities in the first four parentheses are non-zero (which, if not, would presumably indicate either an 'exception' that must be explicitly learned, e.g., 'an historical example' instead of 'a historical example,' or inferior lexicons that need further improvement); we can cancel all of the like terms in their numerators and denominators. Finally, recalling that all of the first factors in these first four parentheses equal $p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)$, and moving the first parenthetical quantity down to be the fourth, gives:

$$[p(\alpha_1\beta_2\gamma_3\delta_4|\varepsilon_5)]^4 = [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\alpha_1\varepsilon_5)]$$
$$\cdot [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\beta_2\varepsilon_5)]$$
$$\cdot [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\gamma_3\varepsilon_5)]$$
$$\cdot [p(\alpha_1\beta_2\gamma_3\delta_4\varepsilon_5)/p(\delta_4\varepsilon_5)]$$
$$\cdot [p(\alpha_1|\varepsilon_5) \cdot p(\beta_2|\varepsilon_5) \cdot p(\gamma_3|\varepsilon_5) \cdot p(\delta_4|\varepsilon_5)],$$

which is the formula used in the report.

