

# Confabulation Theory

## A Synopsis

**Robert Hecht-Nielsen**

*Computational Neurobiology, Institute for Neural Computation, Cal(IT)<sup>2</sup>, ECE Department  
University of California, San Diego, La Jolla, California 92093-0407 USA, r@ucsd.edu*

### Abstract

**A theory of the cognitive function of human cerebral cortex is sketched.**

### 1. Introduction

Confabulation theory (see [1,2] for details beyond this brief sketch) offers a comprehensive, concrete, explanation for cognition. The theory hypothesizes the specific underlying mathematical mechanism of cognition; as well as the neuronal implementation of that mechanism (specified at a ‘meta-level’ of neurophysiological detail: summary descriptions of the dynamical behavior of hypothesized subgroups of cortical neurons).

Confabulation theory proposes that all aspects of cognition (seeing, hearing, command of movement and thought, planning, language, abstract thinking, etc., etc.) are implemented using four fundamental elements: 1) mental object representation, 2) knowledge links, 3) confabulation, and 4) action command origination. These, and their cortical implementations, are briefly sketched, in order, in the following four sections. The concrete numerical values provided in this synopsis of the theory are presented to help fix ideas (many of them probably vary significantly across cortex). If they are within an order of magnitude of being correct I will be happy.

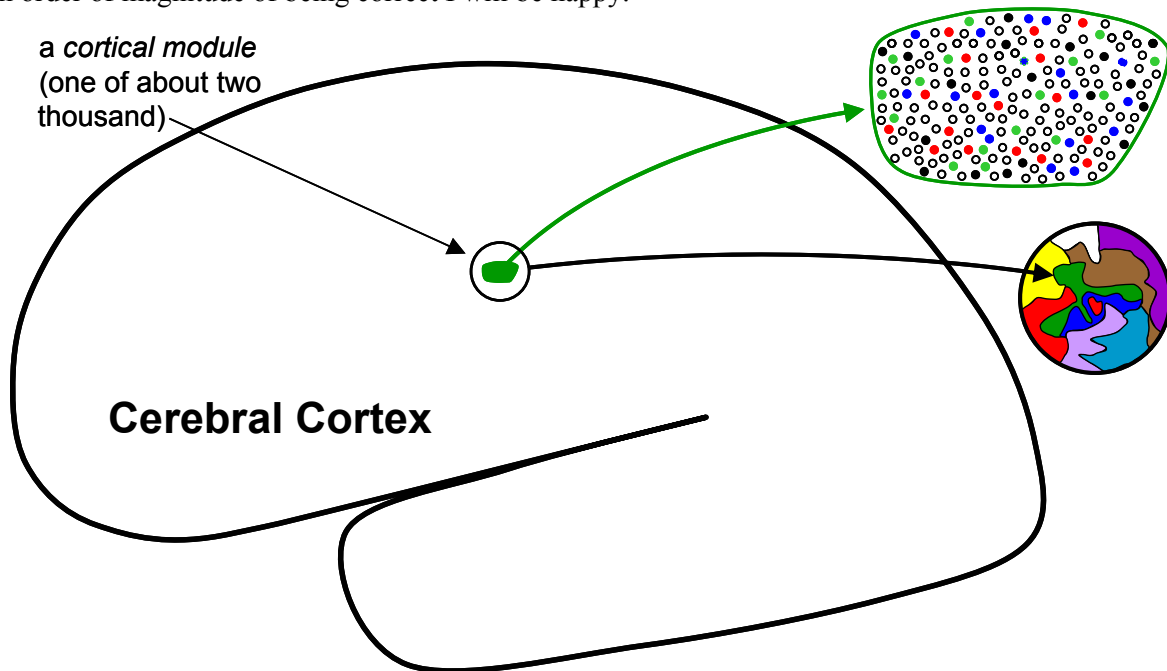


Figure 1. A cortical *module*. Each human cortical hemisphere is comprised of approximately 2,000 of these modules. The upper-right notional, idealized, depiction of the module shows the neuron groups that represent the distinct symbols of the module. The ‘magnified’ depiction beneath illustrates that actual modules are probably irregular in shape. Each module extends through the full depth of cortex and occupies roughly 45 mm<sup>2</sup> of cortical surface area (out of a total of roughly 180,000 mm<sup>2</sup> for both hemispheres).

## 2. Representation of the Objects of the Cognitive World

As illustrated in Figure 1, each hemisphere of human cerebral cortex is hypothesized to be exhaustively divided into roughly 2,000 discrete, localized, largely disjoint and independently controlled, functional *modules*. Each module, extending through the full depth of cortex (i.e., all six Layers) occupies roughly 45 mm<sup>2</sup> of cortical surface area (the average human cerebral cortex possesses roughly 180,000 mm<sup>2</sup> of surface area, including both hemispheres [3]). Anatomical studies of cortical axonal connection patterns [3] have revealed additional fine structure that surely additionally complicates the structure of modules. The exact physical form, and functional details, of cortical modules are not specified by the theory, and are not known. For example, applying Sutton and Strangman's 'network of networks' hypothesis [4], each module could be made up of collections of smaller 'sub-modules.' Another complicating issue is that excitatory cortical neurons having 'pyramidal' morphology (which make up the majority of cortical neurons) almost surely are further divided into many distinct subcategories that serve different functions [5]. Since confabulation theory is new, experimental attempts to delineate, and characterize the functional details of, individual cortical modules in vivo have yet to be launched.

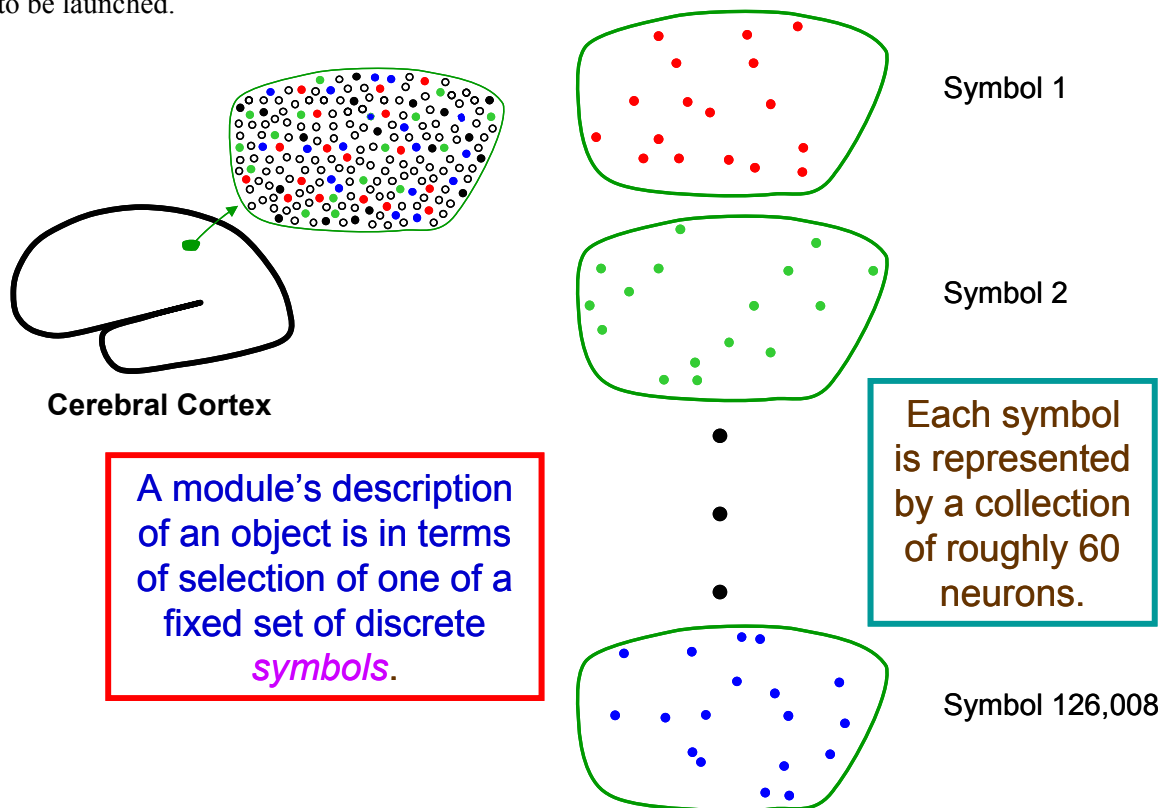


Figure 2. Each cortical module describes a single *attribute* that *objects* of the mental universe may possess. This description, when used, is in terms of selection of a single *symbol* (object attribute descriptor) from among a collection of (typically) thousands of symbols implemented by the module (the particular module shown here is implementing 126,008 symbols). Each symbol is represented by a collection of roughly 60 neurons belonging to a special population of neurons in cortical Layer III of the module. The first key hypothesis of confabulation theory is that this is how the objects of the mental world are represented in the cerebral cortex.

Each cortical module is used to represent one *attribute* that an *object* (visual, auditory, conceptual, abstract, motor process, thought process, plan, etc., etc.) of the cognitive mental universe may possess (see Figure 2). This representation takes the form of the selection of a single *symbol* from among a set of thousands of symbols implemented by the module for describing its object attribute. These symbols are the durable, persistent *terms of reference* for describing the objects of the mental universe. Clearly, such fixed terms of reference must exist if knowledge is to be accumulated over long periods of time.

Confabulation theory postulates that within each cortical module there exists a population of neurons that functions to represent symbols. I believe that these populations reside in cortical Layer III; but the exact location is not important for the purposes of this synopsis. Since each square millimeter of full-depth cortex has roughly 100,000 neurons [3], and this hypothesized population contains roughly 10% of these, each 45 mm<sup>2</sup> cortical module might have about 450,000 of these

symbol representation neurons. If a module implements, for example, 100,000 symbols, and each symbol (for reasons to be indicated in Section 3 below) is represented by a collection of about 60 of these neurons, then it is easy to see that, on average, each symbol representation neuron will participate in representing about 13 different symbols.

The symbolic processing that is the heart of confabulation theory critically depends upon having the symbols of a module be functionally discrete and distinct – not fuzzy and overlapping. Worse yet, as will be discussed in Section 4 below, during confabulation, the symbols must compete with one another on the basis of their representing neuron’s combined total input excitation. Thus, having each symbol representation neuron participate in representing many symbols would seem to be an invitation to ‘crosstalk’ and ‘interference’ between symbols. However, counterintuitively, such problems probably don’t actually arise [1,2].

Cortical modules are organized into *hierarchies*, one symbol in a higher-level module often representing many combinations of symbols at lower levels. [This is an old idea, the power of which has been amply demonstrated by Fukushima’s hierarchical *Neocognitron* family of visual neural networks [6,7,8].] In humans, the modules used to describe language form the core ‘hub’ of cognition. These are connected (via knowledge links) to modules belonging to almost every other functional category of modules.

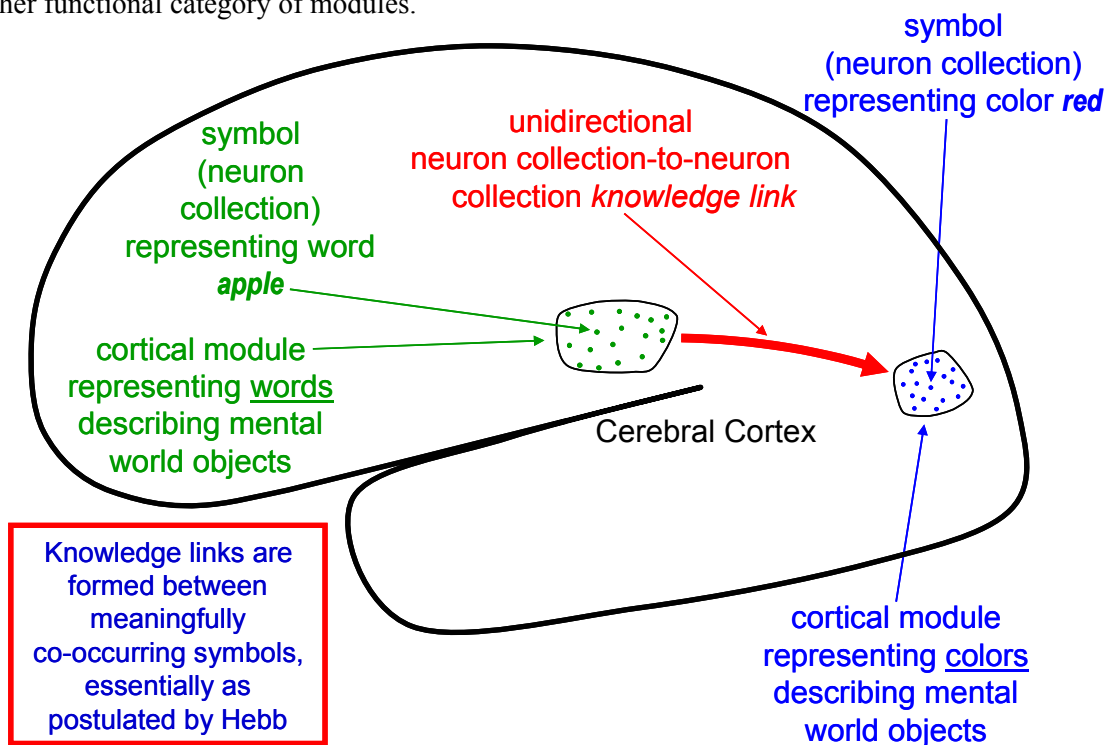


Figure 3. A *knowledge link*. Confabulation theory hypothesizes that all cognitive knowledge is stored in the form of these axonal communication links. Each individual knowledge link is between the collection of neurons representing a particular symbol (termed the *source symbol* of the knowledge link) and members of the collection of neurons representing a second symbol (termed the *target symbol* of the link). As postulated by Hebb in 1949 [9], these links are established on the basis that the involved source and target symbols are meaningfully active at the same time (this is termed *meaningful symbol co-occurrence*). The average human is hypothesized to possess many billions of knowledge links. That knowledge of such a simple kind can explain all of cognition may seem astounding. But that is precisely the second key hypothesis of confabulation theory.

### 3. Cognitive Knowledge

Confabulation theory hypothesizes that all aspects of cognition utilize a simple, uniform type of knowledge: antecedent support axonal links [1,2]. Each such individual *knowledge link* (see Figure 3) connects the neuron collection representing one symbol (termed the *source symbol* of the link) to neurons representing a second symbol (termed the *target symbol* of the link – usually a symbol in a different lexicon from that of the source symbol).

Again, there are details and complications involved. First of all, these axonal links are not direct. They are probably implemented as two-stage Abeles synfire chains [10]. The 60 neurons of the source symbol send axons to millions of neurons scattered all over (many outside its module). Of these, thousands of neurons receive sufficient input from multiple source symbol neurons to become highly *excited*. Thus, this first stage of the chain ‘amplifies’ the high activity of the 60 neurons to high excitation of many thousands of *transponder* neurons (as these intermediate neurons of the chain are termed). Note that the synapses involved in this transponder neuron excitation process will, in general, not be strengthened, since the transponder neurons are typically not already active when the source symbol excitation arrives (thus failing the Hebb co-activity criterion for learning).

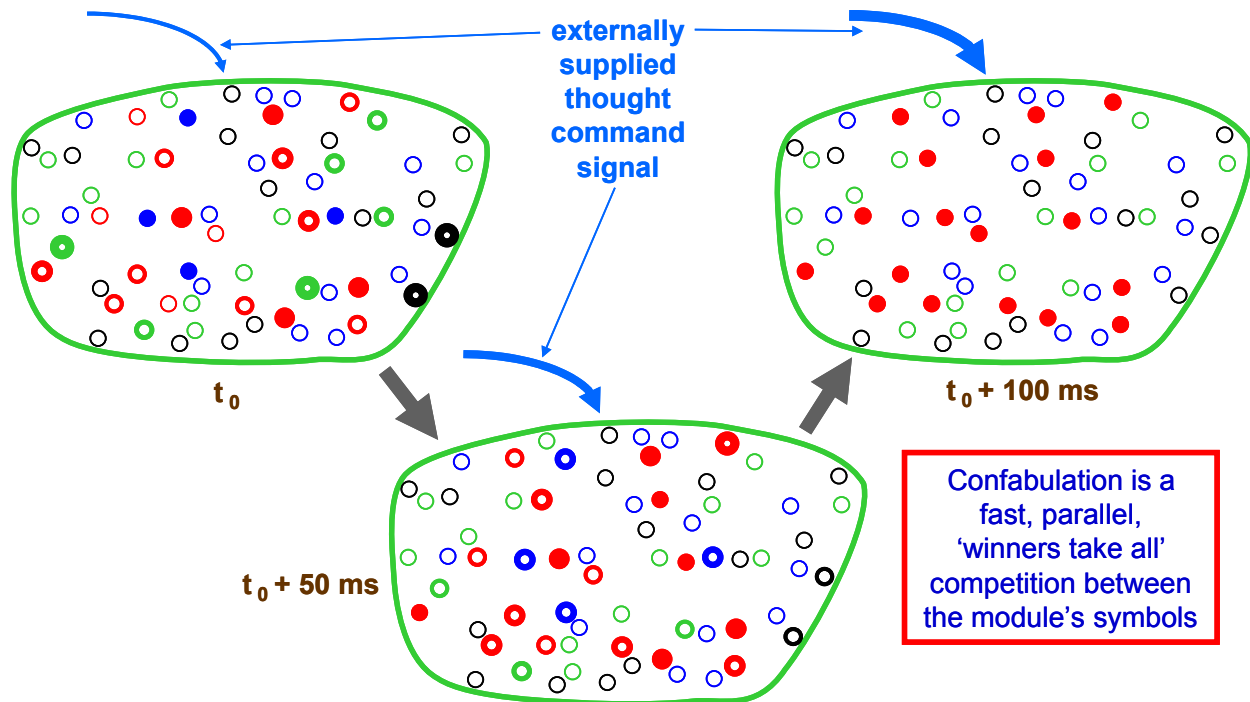


Figure 4. When properly commanded to do so, a module is hypothesized to function as a *neuronal attractor network*. Confabulation theory hypothesizes that all aspects of cognition are carried out using this single ‘information processing operation,’ which is termed *confabulation* (this is the third key hypothesis of confabulation theory). Confabulation is a simple ‘winners take all’ competition process among the symbols of a module. Confabulation takes place only when the module receives a deliberate *thought command signal* (which originates outside the module). The thought command signal is analog (graded), not binary. At the starting time of the confabulation (here denoted by  $t_0$ ), the thought command signal level is low or zero. By rapidly increasing the strength of this command input (which arrives at all points of the module via a small number of parallel axons from an external source which ramify upon entering the module), the symbols compete with one another on the basis of their total excitation at  $t_0$ . The symbol with the highest initial total excitation wins the competition (in this case, the symbol represented by the red neurons). This symbol is termed the *conclusion* of the confabulation. Confabulation is completed in about 100ms. In light of the fact that each module is controlled by a single graded input (just as with the contraction of a muscle), cortical modules can be viewed as the *muscles of thought*.

Of the thousands of transponder neurons that are excited by the momentary source symbol neuron activity, their statistical axonal distributions are assumed to be genetically programmed so that some reasonable fraction (say, 10%) of the neurons representing the target symbol will receive synapses from multiple (perhaps three to six) transponder neurons. During learning, these synapses will be strengthened, since, for learning to take place, the source symbol and the target symbol must be co-active. In other words, these synfire chain links from symbol to symbol will be formed only if Hebb’s co-occurrence condition is met. [New knowledge links are immediately, but temporarily, strengthened when they are first used. Permanent strengthening, if warranted, then occurs over the following few sleep periods [11].]

The average human is postulated to have many billions of individual knowledge links (each of which is termed an *item of knowledge*). This implies a learning rate well in excess of one link per second throughout life. If true, this will have profound implications for our views of human nature, education, etc. For example, a child returning home after a day at school might report that she ‘learned nothing’ that day. In reality, she probably began the process of establishing tens of thousands of new knowledge links. Humans (and other animals) are extremely ‘smart.’

Another implication of this knowledge link hypothesis is that in order for learning to take place ‘on demand’ (i.e., without waiting many days for new, correctly connected, axons to somehow form) there must probably be a vast ‘overwiring’ of cortex. Thus, confabulation theory also proposes that only a small fraction (roughly 1%) of the cortical synapses available for use in storing cognitive knowledge are actually used. Perhaps this is why so many excitatory cortical synapses seem ‘vestigial’ and do not function reliably when tested with patch clamps.

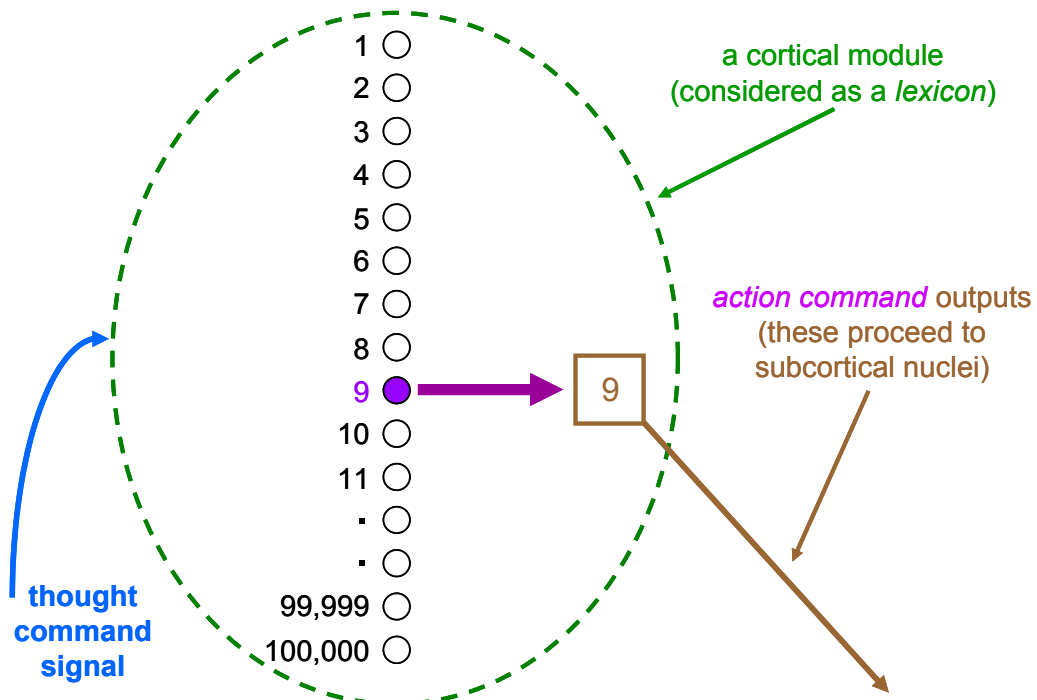


Figure 5. The *conclusion-action principle* (the fourth and last key hypothesis of confabulation theory). Here, a cortical module (illustrated here as an abstract ‘oval’ structure containing a list of symbols to emphasize its function rather than its physiology – in which context a module is referred to as a *lexicon*) has successfully completed a confabulation operation (under control of its externally supplied thought command signal) and reached a conclusion (symbol number 9). Whenever a module reaches a single conclusion it immediately launches a set of *action command* outputs (proceeding to subcortical brain nuclei from neurons in Layer V of the module). The specific action command outputs that are launched are those which have (via a learning process not discussed here – see [20,21] for more insight) been previously *associated from* that specific conclusion symbol. These action command outputs can cause behaviors to occur. The conclusion-action principle is hypothesized to be the origin of all non-autonomic and non-reflexive behavior. During wakefulness, many behaviors (most of them small ‘microbehaviors’) are launched every second.

#### 4. Confabulation

Besides implementing a lexicon of symbols for describing that module’s mental object attribute; each cortical module is also responsible for carrying out the *confabulation* operation (a ‘winners take all’ competition process among the symbols of the module – see Figure 4). Confabulation is postulated to be the only information processing operation used in cognition. The hypothesized neuronal *attractor network* [12,13,14,15,16,17] mechanism by which modules (when deliberately commanded to do so by a single, deliberate, *thought control signal* input to the module from an external source) carry out confabulation is illustrated in Figure 4.

Figure 4 shows many symbols of the module (each notionally represented by colored circles having thicknesses directly related to their excitation levels) receiving various levels of input excitation at a starting time  $t_0$  from incoming knowledge links (as shown, each such knowledge link typically only effects a subset of each involved target symbol’s neurons). Delivery of a thought control signal input (illustrated in blue) of rapidly increasing amplitude (the arrows above the module in the figure) to the module then causes that symbol with the highest average level of input excitation to end up having all 60 of its neurons highly active (for a brief instant); with all other symbol representation neurons of the module silent. This symbol is termed the *conclusion* of the confabulation (confabulations can also end with multiple excited symbols or no symbol; but this complication is not discussed here). Confabulation involves a multitude of parallel, local interactions between the involved neurons during the 100 ms required to complete the process.

That such a simple ‘winner take all’ process is able to account for all aspects of cognition is a monumental claim. But that is exactly what confabulation theory postulates.

In the confabulation theory view of cortical cognitive function (cortex does other things as well, besides cognition, such as triggering *behaviors* – see next section), each module has its cognitive activity controlled by a small number of thought control input afferents from outside the module. Thus, the theory views a cortical module as an exact analog of a muscle: a discrete ‘action unit’ controlled by a single graded input. The exact origin of the thought control axonal inputs to cortical modules is not known; but I suspect they arise in one or more subcortical nuclei (see [18,19] for intriguing clues). In summary, confabulation theory views the cortical modules as the *muscles of thought*.

## 5. The Origin of Behavior

As illustrated in Figure 5, confabulation theory hypothesizes that every time a confabulation operation carried out by any cortical module yields a *definitive conclusion* (namely, one symbol – not multiple symbols or no symbol), a set of *action* (movement process and/or thought process) *commands* associated from that particular conclusion symbol are immediately launched. This explains the continual flow of behaviors that emerge, moment by moment, during wakefulness. In effect, each new behavior is a response to the latest change in the representation of the mental world state. Action commands originate in Layer V of cortex and typically target subcortical nuclei such as motor or thought nuclei or the basal ganglia.

The implemented action commands produced by this *conclusion-action principle* of the theory are termed *behaviors*. Most behaviors are small ‘housekeeping’ functions (e.g., the next small segment of a movement or thought process); which are termed *microbehaviors*. Tens of microbehaviors are often implemented in one second. Higher-level behaviors (created by conclusions on higher-level, more abstract modules – often residing in frontal cortex), such as a decision to take a trip to Copenhagen, are launched much less frequently.

In effect, behavior results (during wakefulness) each time the state description of the mental world is updated (by activation of a new confabulation conclusion symbol in a module). Each successful confabulation launches the next behavior, and so on, endlessly until the next sleep period. The wizard homunculus standing behind the curtain pulling the levers of behavior is thereby exorcised. All non-reflexive and non-autonomic behavior is postulated to originate in this way.

## 6. The Mathematics of Cognition

That the four key elements of confabulation theory, described in the previous four sections, are capable of explaining every aspect of cognition may seem astonishing. In this section, the underlying mathematics of confabulation is briefly discussed to see why this assertion may be tenable.

Confabulation theory proposes that the underlying mathematical process of cognition is *maximization of cogency* (see [1] for more details). For example, if we consider four *assumed fact* symbols:  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  (these are symbols being expressed on four different modules; with each symbol transmitting excitation, via all the available knowledge links, to target symbols of a fifth *answer* module that is about to undergo confabulation).

Confabulation theory hypothesizes that the fundamental underlying mathematical operation of cognition is to find that symbol  $\epsilon$  of the answer module which maximizes *cogency*  $p(\alpha\beta\gamma\delta|\epsilon)$ . Cogency is the probability of the assumed facts being true, given an assumption that the symbol  $\epsilon$  is true. In other words, confabulation theory claims that each decisionmaking process involved in cognition is selection of that conclusion which is most supportive of the assumed facts actually being true.

An important (rigorously proven [1]) property of cogency maximization is that in a ‘logical information environment’ (playing chess, doing mathematics, etc.) it yields the same results as classical Aristotelian logic. Thus, the cogency maximization hypothesis implies that cognition is ‘logical’ when that is possible (most of the information environments we encounter are not logical). In a ‘non-logical’ environment (e.g., when parking your car), cogency maximization just picks the conclusion that best supports the probability of the available facts being true; and then moves on (there is no ‘logical guarantee’ that the conclusions reached are correct, but that is not important). Interestingly, cogency maximization is different from, and inconsistent with, maximization of a posteriori probability (selecting the conclusion

which has the highest probability of being correct, given the assumed facts); which is a popular, but egregiously wrong [1], model of cognition.

As shown by the **Fundamental Theorem of Vertebrate Cognition** [1], confabulation can, under mathematical conditions which confabulation theory postulates animal neurological evolution has been able to satisfy, approximate cogency maximization. This is important, because cogency maximization itself cannot be carried out in practice.

In cognition, confabulations are rarely carried out alone. They normally occur in contemporaneous, mutually interacting, groups. This adds greatly to the information processing ‘value’ that can be achieved. For example, during such multiple confabulations, the excitation of different symbols in each module can change dynamically; altering the outcome of the winner take all competition processing going on in that module. Tens of millions of constraints on the ‘solution’ (in the form of knowledge links between the involved candidate conclusions being considered in the various involved modules) can be applied in parallel; with the highest-cogency set of conclusions ‘bubbling to the surface’ at the end of the process (which is often completed in a tenth of a second). To succeed in converging to such a *confabulation consensus* of conclusions requires that the analog thought control signals be properly manipulated in a dynamic, coordinated manner; exactly as the motorneuron inputs to muscles must be properly manipulated in a dynamic, coordinated manner for a movement to be successfully executed. Confabulation theory thus views thinking and movement as very similar.

In summary, confabulation can be thought of as a practically implementable decisionmaking tool for finding the ‘best’ conclusion to a very general type of question. This accounts for its universal applicability across all the various functional domains of cognition.

Confabulation theory also involves the intimate relationship between cortex and thalamus (which has not even been mentioned herein – see [22] for some insights). There is much more to say, but hopefully this synopsis has provided a rough idea of the theory’s content.

## 7. Technological Application of Confabulation Theory – i.e., AI

Beyond its potential scientific utility as an explanation of human cognition, confabulation theory also has implications for technology. Specifically, it is relatively straightforward to implement *confabulation architectures* consisting of lexicons and knowledge links on a computer.

There are three main challenges in implementing computer cognition using a confabulation architecture:

1. Conceiving and then precisely defining a confabulation architecture (lexicons and knowledge bases) that will be capable of carrying out the desired cognitive function(s).
2. Conceiving and then precisely defining the thought processes (spatiotemporal collections of thought commands – usually divided up into small, fixed, action command chunks to be associated from, and triggered by, specific confabulation conclusion symbols) that will be used to control the function of the confabulation architecture (and of any other attached architecture – such as mechanical and sensory effectors of various kinds).
3. Conceiving and executing an appropriately staged sequence of *learning opportunities*, during which the knowledge links of the knowledge bases will be progressively created; where, typically, each learning opportunity requires exposure to some sort of carefully prepared external information environment.

Obviously, applying confabulation theory involves issues, techniques, and design methodologies that are utterly alien in comparison with today’s information technology. Creating these capabilities will likely require the expenditure of vast resources. How can we gain confidence in advance that such expenditures will pay off?

One way is to start with ‘baby applications’ of confabulation theory that, while challenging, can be carried out by a small group of researchers working over a few years. Well, this has been done. And the results indeed seem encouraging. For examples, see [1,2].

## References

1. R. Hecht-Nielsen (2005) Cogent confabulation. *Neural Networks* **18**:111--114.
2. R. Hecht-Nielsen **Mechanization of cognition** in: Y. Bar-Cohen [Ed] **Biomimetics** [CRC Press, Boca Raton, FL 2005].
3. G. Paxinos & J. K. Mai [Eds] **The Human Nervous System, Second Edition** [Academic Press, San Diego 2003].
4. J. P. Sutton & G. Strangman **The Behaving Human Neocortex as a Dynamic Network of Networks** in: R. Hecht-Nielsen & T. McKenna [Eds] **Computational Models for Neuroscience** [Springer, London 2003].
5. H. Markram **Elementary Principles of Nonlinear Synaptic Transmission** in: R. Hecht-Nielsen & T. McKenna [Eds] **Computational Models for Neuroscience** [Springer, London 2003].
6. K. Fukushima (1975) Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics* **20**: 121–136.
7. K. Fukushima, S. Miyake & T. Ito (1983) Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-13**:826–834
8. K. Fukushima (2005) Restoring partly occluded patterns: a neural network model. *Neural Networks* **18**:33–43.
9. D. Hebb **The Organization of Behavior** [Wiley, New York 1949].
10. M. Abeles **Corticonics** [Cambridge University Press, Cambridge 1991].
11. T. J. Sejnowski & A. Destexhe (2000) Why do we sleep? *Brain Research* **886**:208–223.
12. J. A. Anderson, J.W. Silverstein, S.A. Ritz & R.S. Jones (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review* **84**:413–451.
13. S.-I. Amari (1974) A method of statistical neurodynamics. *Biological Cybernetics* **14**:201–215.
14. J.J. Hopfield (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings National Academy Science* **79**:2554–2558.
15. M.A. Cohen & S. Grossberg (1983) Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions Systems Man & Cybernetics* **13**:815–826.
16. K. Haines & R. Hecht-Nielsen (1988) A BAM with increase information storage capacity. *Proceedings, 1988 International Conf. on Neural Networks* [IEEE Press, Piscataway NJ] **I**:181–190.
17. D. Amit **Modeling Brain Function: The World of Attractor Networks** [Cambridge University Press, Cambridge 1989].
18. S. Herculano-Houzel, M.H.J. Munk, S. Neuenschwander, W. Singer (1999) Precisely synchronized oscillatory firing patterns require electroencephalographic activation. *Journal of Neuroscience* **19**:3992–4010.
19. P. Fries, J.H. Reynolds, A.E. Rorie, R. Desimone (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**:1560–1563.
20. J.W. Brown, D. Bullock & S. Grossberg (2004) How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks* **17**:471–510.
21. T. Shibata, T. Tabata, S. Schaal & M. Kawato (2005) A model of smooth pursuit in primates based on learning the target dynamics. *Neural Networks* **18**:213–224.
22. M.S. Sherman & R.W. Guillery **Exploring the Thalamus** [Academic Press, San Diego 2001].