How Might the Brain Represent Complex Symbolic Knowledge?

Ben Goertzel

Abstract— **A novel category of theories is proposed, providing a potential explanation for the representation of complex knowledge in the human (and, more generally, mammalian) brain. Firstly, a "glocal" representation for concepts is suggested, involving localized representations in a sparse network of "concept neurons" in the Medial Temporal Lobe, coupled with a complex dynamical attractor representation in other parts of cortex. Secondly, it is hypothesized that a combinatory logic like representation is used to encode abstract relationships without explicit use of variable bindings, perhaps using systematic asynchronization among concept neurons to indicate an analogue of the combinatory-logic operation of function application. While unraveling the specifics of the brain's knowledge representation mechanisms will require data beyond what is currently available, the approach presented here provides a class of possibilities that is neurally plausible and bridges the gap between neurophysiological realities and mathematical and computer science concepts.**

I. Introduction

THE human mind is able to represent, generate and manipulate complex relational knowledge, such as this sentence. According to our current, standard models of neurodynamics, however, there is no clear explanation for how this sort of complex relational knowledge might emerge from brain matter.

The lack of detailed neural theories in this regard is certainly understandable, because available neurological data is not obviously adequate to validate or refute such theories. If brain imaging tools with high simultaneous spatial and temporal resolution were available, we could make a much more straightforward effort at exploring this aspect of the brain empirically.

On the other hand, sometimes creative experiments are driven by the existence of specific theories in need of validation or refutation. With this in mind, it seems worthwhile to explore possible theoretical explanations that might help us understand how systems like brains could give rise to the ability to create and manipulate complex, abstract formal systems. Here we present one possible theory in this direction.

The theory suggested here synthesizes recent neuroscience findings from various directions, including concept neurons , dynamical attractors and temporal phase coherence, with classical mathematical notions such as combinatory logic. While it is unlikely that the precise details outlined here will turn out to be the way the brain works, we present a somewhat specific theory here largely as a placeholder indicating a certain *kind* of theory. Something like this, we suggest, is reasonably likely to be the way the brain does it.

Ben Goertzel (email: ben@goertzel.org) is with the OpenCog Foundation, 35516 Aspen Court, Rehoboth DE, 19971, USA

Specifically, the hypothesis we explore here is that:

1) The brain represents complex relationships in the style of illative combinatory logic, in which variables don't need to be explicitly represented. In place of variables one has "higher order functions", i.e. functions that transform other functions. In place of traditional logical quantifiers one has an XOR type operator.
2) "Concept neurons" in the Medial Temporal Lobe (MTL) serve as "symbols" for cortical activity patterns, in the sense that individual concept neurons correspond to particular distributed attractors in the cortex
3) The instruction for one cortical attractor F to transform another cortical attractor G, is encoded in a relationship of *systematically asynchronous firing* [1] between the concept neuron c(F) corresponding to F and the concept neuron c(G) corresponding to G. That is, a brief delay between c(G) firing and c(F) firing, encodes the function application F(G)
4) *Relational grouping* is implemented via *temporal grouping*. That is: the disambiguation between e.g. $F(GH)$ and $(FG)H$ is implemented via, in effect, having a longer firing gap between two symbols separated by a parenthese.

This is an admittedly speculative model. However, it is reasonably concrete, and utilizes only mechanisms with known neurological plausibility.

II. The Binding Problem

The issue of the neural representation of complex relational, symbolic knowledge, which we consider here is closely tied to the so-called 'binding problem," one of the most enduringly vexing problems in cognitive science. Feldman [2] has decomposed the binding problem into several subproblems:

1) Coordination among disparate brain regions involved in the same activity (e.g. temporal coherence). .
2) The subjective unity of perception
3) Visual feature-binding
4) Variable Binding

The first of these, coordination, is moderately well understood at present. There are mechanisms like phase synchronization with the capability to dynamically bind far-flung neural subnetworks into coherent activity. There are also specific neural circuits and connectivity patterns that encourage coordination, which sometimes get short shrift in the literature with all the focus on phase synchronization

Regarding the subjective unity of perception, one thing that has become clear as we have understood the brain more fully in the last half-century, is that there is no specific location in

the brain where, say, the whole perceived visual (or auditory, or tactile, or multi-sensory) "scene" of the currently present world is stored. The problem of how a physically disparate collection of neural subnetworks contribute to produce a subjectively coherent, felt scene, is a sub-case of the overall problem of consciousness and qualia [3]. If one accepts that, as Greenfield [4], Christos [5] and many others have proposed, consciousness is associated with distributed attractors in the brain, then the subjective unity of perception becomes less mysterious, and is decomposed into

- the neurodynamical problem of how attractors emerge in the brain
- the philosophical "hard problem" of qualia

Freeman [6], [7], [8] and others have given detailed explanations of the emergence of attractors in the brain, using mathematical, computer-simulation, and neurophysiology-driven arguments. There is clear evidence for the role of attractors in some elements of brain function (e.g. olfaction) and less clear in others; and gathering data about such matters is difficult due to the lack of any brain scanning technology combining high spatial and temporal resolution.

The problem of qualia, on the other hand, does not clearly lie wholly within the scope of science as currently understood. In prior publications [9] we have addressed qualia from a panpsychist perspective, which appears to eliminate any logical contradictions associated with qualia, but is sometimes perceived as counterintuitive. In any case we will not explore this aspect further here.

Finally, variable binding is generally recognized as the most difficult aspect of the overall "binding problem." Prototypical examples involve relational reasoning such as

$$owns(z, y)\&gives(z, x, y) \rightarrow owns(x, y)$$

A modest-sized literature has arisen, focused on the conception of mechanisms via which variable binding might be realized in "connectionist" systems with purely distributed representations. van der Velde and de Kamps [10] survey much of the literature as of 2006.

Temporal phase synchrony plays a large role in many of these theories, such as Shastri's influential work with SHRUTI [11], [12]. For instance, it has commonly been proposed that when the neuron or neural subnetwork representing a function and the neuron or neural subnetwork representing an argument are firing in-phase together, this may connote the function taking the argument as an input. This kind of phase coherence is relatively slow, so seems not that likely to be the key mechanism binding together, say, the disparate visual features of a scene into a coherent perceived whole. However, variable binding is arguably a slower process, hence phase coherence remains a plausible candidate here.

A related, alternative approach posits systematic asychrony rather than synchrony as the basis of binding between a function and its argument. Love [1] points out that the function-argument relationship is basically asymmetric, so that it might make sense to posit that the neural representation of $f(x)$ involves the f neural subnetwork firing systematically slightly *before* the x neural subnetwork.

The work of Hummel et al [13], which has been oriented largely toward understanding the neural basis of relational inference in the context of analogy reasoning, also relies on the notion of temporal phase relationships in a broad sense, though it could arguably be reconciled with either synchrony or systematic asynchrony based accounts.

Signature propagation approaches like that of Browne and Sun [14], on the other hand, suggest that each variable in an expression corresponds to its own neural group; and each concept to which a variable may be bound, corresponds to a particular "signature" that a neural group can output. In essence, a signature serves as a name for a concept. A shortcoming of this approach is that, in its simplest form, it requires a unique signature for every representable object so a new signature must be created for each new item encountered. There are modified versions of the approach in which signatures are allocated dynamically only to those concepts being currently thought about – so that the signature used for "chair" at one moment, might be re-used for "aardvark" later, depending on which concepts are currently being focused on and hence in need of signatures [15]. Overall, this approach feels very computer-sciency and has not been closely tied to neurological structures or detailed observations about brain dynamics.

These various theories are interesting and likely have value in various ways. They point to various ingredients of neural representation and behavior. However, recent discoveries regarding the occurrence of localized representation in the brain would seem to suggest that the various lines of thinking regarding the neural foundations of variable binding proposed in the literature should be revisited, reinterpreted and perhaps heavily revised.

III. Glocal Memory in the Brain

The naive notion that the brain stores its knowledge like a semantic network – with neurons or localized neuronal groups representing individual concepts, and related concepts interlinked by bundles of synapses – was refuted long ago. Classic examples of "holographic" memory storage in the brain point out that sometimes the brain goes to the opposite extreme – knowledge can be stored across a wide area of the brain, so that when any chunk of that brain region is removed, the knowledge still remains, at least in an approximative form. This sort of distributed memory can be modeled mathematically using Hopfield type associative memory networks [16], [17], and studied using methods of nonlinear dynamics.

On the other hand, the "semantic network" type model of neural knowledge representation has more meat to it than early advocates of distributed, holographic neural memory realized. The notion of a "grandmother cell" – a neuron representing an individual concept like one's grandmother – was dismissed and even mocked by a subset of the scientific community for quite some time, yet has recently been definitively resurrected via striking experimental results.

It is now clear that the human brain – in particular, the MTL – *does* possess individual cells that respond differentially to very particular concepts [18], [19] – though the literature has focused on neurons firing differentially in response to the actresses Jennifer Aniston and Halle Berry, rather than grandmothers per se [1]. The precise extent and importance of this "concept cell" phenomenon remains to be determined, but at very least this is a highly thought provoking development. It seems unlikely there is a *single* Jennifer Aniston cell exists in anyone's brain; rather, it seems that a concept like this is represented by a very sparse, distributed network of "concept neurons." Exactly how many Jennifer Aniston neurons an individual's brain is likely to contain remains unknown – "thousands" might be a reasonable guess at present.

However, the existence of concept cells in the MTL does not obviate the importance of distributed representations in the brain. For one thing, obviously, if there are 1000 Jennifer Aniston neurons that somehow synchronize together, then we still have a distributed representation, just a sparse one. Secondly, it seems clear that these concept neurons are not the whole story. While the whole story of neural knowledge representation remains largely unknown, one highly plausible perspective, given all the data available, is that localized representations in MTL are coupled and coordinated with distributed, nonlinear attractor based representations in the cortex. This would be an instance of what I have called "glocal memory" [20] – a term describing memory structures in which each memory item stored has both a distributed ("map") aspect and a localized ("key") aspect. Glocal memory has certain advantages in an AI context, because it allows an AI system to exploit the facility of local representations for explicit symbolic reasoning, alongside the facility of distributed Hopfield-net style representations for associative memory and creative concept-generation. It may provide brains with similar advantages.

IV. Combinatory Logic

Combinatory logic started with a paper by Moses Schonfinkel [21], written with the aim of figuring out how to do logic without bound variables. In one of the ultimate acts of mathematical reduction, he reduced logic to a simple language consisting of one constructor (function application) and some primitive constants. This work was continued by Haskell Curry [22] who introduced modern combinatory logic notation and developed a body of related theory. At about the same time, Alonzo Church introduced the lambda-calculus as a new way to study the concept of rule. Originally his purpose was to provide a foundation for mathematics. Combinatory logic and lambda-calculus, in their type-free versions, generate basically the same algebraic and logic structures. Today combinatory logic lives on primarily among those developing and theorizing about functional programming languages such as Haskell.

[1]In time these actresses may become grandmothers themselves, providing the neuroscience literature with greater metaphorical coherence.

The notation of combinatory logic relies on the notion of currying, in which adjacency indicates function application and binds to the left, so that e.g.

$$Sfgx \equiv ((S(f))(g))(x)$$

One introduces a set of combinators, and then expresses general logical relationships via combining them in complex expressions. Some standard combinators are:

```
S f g x => ( f x ) ( g x )
K x y   => x
I x     => x
B f g x => f ( g x )
C f x y => f y x
W f x   => f x x
D f     => f f
Y f => Y ( Y f )
```

It is a well known, simply proved theorem that the S and K combinators form a complete set, and can be used to generate any other combinator. Schonfinkel also gave a single combinator which possesses this completeness property. The Y combinator is famous as the simplest archetypal form of recursion.

Theoretical computer scientists should note that, by adding an extensionality rule to combinatory logic – i.e. $\forall x \{(F\,x) = (G\,x)\} \Rightarrow F = G$ – one obtains an equational theory that corresponds exactly to $\beta\eta$-equivalence. "Illative combinatory logic" involves the augmentation of the basic combinators with simple rules such as this, enabling combinatory logic to do everything that ordinary formulations of logic do, but without explicit manipulation of variables.

A. An Example of Variable Elimination

For a concrete example of how combinators enable variable-free expression of symbolic structures, let us turn to the OpenCog AI framework [23], [24], whose knowledge representation features (among many other things) logical relations such as

```
AND
   InheritanceLink $X cat
   eats $X mice
```

This would be expressed in a more standard notation as

```
inheritance($X, cat) & eats($X, mice)
```

However one phrases it, though, this example involves the variable $X. How can we get rid of the $X?

The easiest route here involves the C combinator, defined above by

```
C f x y = f y x
```

Using this tool,

```
InheritanceLink $X cat
```

becomes

```
C InheritanceLink cat $X
```

and

```
(eats $X) mice
```

becomes

```
C (eats) mice $X
```

so that overall we have

```
AND
   C InheritanceLink cat
   C eats mice
```

where the C combinators essentially give instructions as to where the "virtual argument" X should go.

In this case the variable-free representation is basically just as simple as the variable-based representation. This won't always be the case – sometimes, in a computer science context, execution efficiency will be significantly enhanced by use of variables.

Conceptually, though, the elimination of variables provides a dramatic simplification. One no longer has the peculiar constructs of variables that need to be assigned to values. The "variable binding problem," as such, simply goes away! In the place of variables bound to values, what one gets in return are higher-order functions – functions that take other functions as arguments, and so on.

B. Neural Realization of Combinators

How might combinators, or something like them, be expressed in the brain? Suppose that mathematical functions are to be represented as neural subnetworks. Then, the main "trick" that needs to be carried out is to have neural subnetworks manipulate *each other*. Note that this is fundamentally different than having one neural subnetwork pipe its output into another neural subnetwork. Rather, in some sense, neural subnetwork A must act on neural subnetwork B, producing neural subnetwork C as the result of this activity.

To see how this formal notion might be relatively simply instantiated in the brain, consider a neural subnetwork which contains a subset of neurons interpretable as "control parameters", and then other neurons serving as input, output and internal state. The state of the control parameters is viewed as determining which mathematical function the subnetwork computes. In this framework, what would it mean for neurally-implemented function A to act on neurally-implemented function B, producing neurally-implemented function C as output? One possible meaning would be for a subnetwork computing A to spread activity into the control parameters of a subnetwork computing B, thus causing this latter subnetwork to change and start computing C instead.

Note that the same brain could potentially contain many different subnetworks redundantly computing B, so that the transformation of one of these into a subnetwork computing C instead is no great loss. This kind of redundancy doesn't seem difficult to achieve neurally; the same activation pattern needed to cause one subnetwork to start computing some particular function, could also be sent to other subnetworks simultaneously, causing them to start computing the same function.

So, for instance, a neural subnetwork playing the role of the C combinator, could act on the control parameter subset of a neural subnetwork playing the role of "eats", transforming the latter into a subnetwork that carries out a certain other transformation T involving "eats" on other subnetworks. In particular, this other transformation T would transform a subnetwork Y into a subnetwork Z that would map its argument X into : the output of applying the result of "eats" to Y, to X.

Yes, this looks a bit convoluted when you write it down in English. But so would the spreading of activation in any reasonably complicated recurrent neural network. The point is that all this complication can, mathematically, be instantiated by neural subnetworks spreading activation into each other's control-parameter subsets, and taking each others' control-parameter subnetworks as inputs. Neurons acting as control parameters for subnetworks, together with control-parameter neurons acting as inputs to neural subnetworks, enable sets of neurons to act implicitly as functions acting on functions on neural subnetworks, functions acting on functions acting on functions on neural subnetworks, and so forth.

V. An Hypothesis Regarding the Representation of Complex Symbolic Knowledge in the Brain

Now we have articulated all the ingredients needed to posit a fairly concrete hypothesis regarding the neural representation of complex symbolic knowledge.

Firstly, the variable binding problem is bypassed, via hypothesizing that the brain represents complex relationships in the style of illative combinatory logic, in which variables don't need to be explicitly represented. In place of variables one has "higher order functions", i.e. functions that transform other functions. Eliminating bound variables doesn't, in itself, solve the problem of neural-symbolic knowledge representation, but it casts the problem in a more tractable form.

The actual transformations indicated via combinatory logic are proposed to be carried out by neural subnetworks that "act on each other" via modifying neurons within each subnetwork that play the role of control parameters (determining what mathematical function the subnetwork computes). These neural subnetworks will generally be physically distributed throughout the brain, and are likely to maintain their activity over time due to complex "strange attractor" and "strange transient" style dynamics.

"Concept neurons" in MTL are suggested to serve as "symbols" for cortical activity patterns, in the sense that individual concept neurons correspond to particular distributed attractors in the cortex. Patterns of synaptic connectivity between concept neurons are hypothesized to encode instructions for actual transformations enacted by corresponding cortical attractors upon each other.

Finally, we suggest that the instruction for one cortical attractor F to transform another cortical attractor G, is encoded in a relationship of *systematically asynchronous*

firing [1] between the concept neuron c(F) corresponding to F and the concept neuron c(G) corresponding to G. That is, a brief delay between c(G) firing and c(F) firing, encodes the function application F(G).

The question then arises: How does the brain disambiguate between e.g. $F(GH)$ and $(FG)H$? One obvious possibility is that this is implemented via, in effect, having a different firing gap associated with each parenthetical grouping. In other words: the logical grouping denoted syntactically by parentheses, in the standard combinatory logic notation, might be represented in the brain via temporal grouping. This would mean that in an expression involving 10 levels of nested parentheses, 10 different gradations of timing would be required. However, it's clear that human brains are not capable of arbitrarily complex logical manipulations, without aid of external devices like paper and pencil or computers. So the fact that a certain mechanism would become increasingly awkward as the logical expressions involved become more complicated, seems not a counterargument against neural or psychological plausibility.

This is an admittedly speculative model. However, it is reasonably concrete, and utilizes only mechanisms with known neurological plausibility. There seems little likelihood that the brain handles symbolic knowledge *exactly* as proposed here, with no variations or added complications. However, there also seems no good reason, based on the available data, why *something like this* couldn't be the true story.

VI. Conclusion

Detailed understanding of how the brain realizes complex symbolic representations and manipulations remains for the future, when brain imaging has advanced further. Whether this future will come in years or decades remains to be seen. In the meantime, however, we can seek to understand what *kind* of theory might appropriately bridge the gap between neurons and symbols. Here I have presented one relatively concrete hypothesis in this regard.

In this hypothesis, combinatory logic is used to bypass the variable binding problem. Concept neurons are used to avoid the need to make all representations purely distributed, but are proposed to work together with distributed attractor representations. Concept neurons and their interrelationships are posited to encode combinatory logic style representations, while the correlated attractor subnetworks actually execute the transformations implied by the combinators. Systematic asynchrony is hypothesized as a tool for neural realization of function applications, including parenthetically nested ones.

These various notions, put together, provide a plausible explanation how a system like a brain could plausibly give rise to complex symbolic structures such as one sees in logic and mathematics, and such as humans appear to utilize when carrying out deliberative reasoning about their everyday lives. The next challenges along the path blazed here would be to implement the ideas suggested in a computational simulation, and to seek explicit neuroscience evidence in favor of the ideas given.

References

[1] B. Love, "Utilizing time: Asynchronous binding." *Advances in Neural Information Processing Systems*, vol. 11, pp. 38–44, 1999.

[2] J. Feldman, "The neural binding problem(s)," *Cognitive Neurodynamics 7*, 2013.

[3] T. Metzinger, *Neural Correlates of Consciousness*. Bradford, 2000.

[4] G. SA and C. T. F, "A neuroscientific approach to consciousness." *Prog Brain Res.*, 2005.

[5] G. Christos, *Memory and Dreams: The Creative Human Mind*. Rutgers University Press, 2003.

[6] G. Li, Z. Lou, L. Wang, X. Li, and W. J. Freeman, "Application of chaotic neural model based on olfactory system on pattern recognition," *ICNC*, vol. 1, pp. 378–381, 2005.

[7] W. Freeman, *Societies of Brains*. Erlbaum, 1995.

[8] R. Kozma, M. Puljic, and W. Freeman, "Thermodynamic model of criticality in the cortex based on eeg/ecog data," in *Criticality in Neural Systems , Ed. by Plenz, D. and Niebur, E.* Wiley, 2013.

[9] B. Goertzel, *The Hidden Pattern*. Brown Walker, 2006.

[10] F. van der Velde and M. de Kamps, "Neural blackboard architectures of combinatorial structures in cognition." *Behavioral and Brain Sciences*, vol. 29, pp. 37–70, 2006.

[11] L. Shastri and V. Ajjanagadde, "From simple associations to systematic reasoning: A connectionist encoding of rules, variables, and dynamic bindings using temporal synchrony," *Behavioral & Brain Sciences*, vol. 16-3, 1993.

[12] L. Shastri, "Episodic memory and cortico-hippocampal interactions." *Trends in Cognitive Science*, 2002.

[13] J. Hummel, "Getting symbols out of a neural architecture." *Connection Science*, 2011.

[14] S. R. Browne A, *Connectionist variable binding.* Springer, 2000.

[15] L. B. andJerome Feldman and L. M. Dermed, "A (somewhat) new solution to the binding problem," *Neural Computation*, pp. 2361–78, 2008.

[16] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. of the National Academy of Sciences*, vol. 79, pp. 2554–2558, 1982.

[17] D. J. Amit, *Modeling brain function – the world of attractor neural networks*. Cambridge University Press, New York, USA, 1989.

[18] R. Q. Quiroga, L. R. amd G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single-neurons in the human brain." *Nature*, vol. 435, pp. 1102–1107, 2005.

[19] R. Q. Quiroga, "Concept cells: The building blocks of declarative memory functions," *Nature Reviews Neuroscience*, vol. 13, pp. 587–597, 2012.

[20] B. Goertzel, J. Pitt, M. Ikle, C. Pennachin, and R. Liu, "Glocal memory: a design principle for artificial brains and minds," *Neurocomputing*, Apr. 2010.

[21] M. Schonfinkel and P. Bernays, "Zum entscheidungsproblem der mathematischen logik," *Mathematische Annalen*, p. 99:34272, 1929.

[22] H. B. Curry and R. Feys, *Combinatory Logic*. Amsterdam, Holland: North-Holland, 1958.

[23] B. Goertzel, C. Pennachin, and N. Geisweiller, *Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy*. Springer: Atlantis Thinking Machines, 2013.

[24] ——, *Engineering General Intelligence, Part 2: The CogPrime Architecture for Integrative, Embodied AGI*. Springer: Atlantis Thinking Machines, 2013.