Neural Algebra and Consciousness: A Theory of Structural Functionality in Neural Nets

Erwin Engeler

Department of Mathematics,
Federal Institute of Technology,
HUT E 31, 8092 Zurich, Switzerland
engeler@math.ethz.ch
http://www.math.ethz.ch/~engeler

Abstract. Thoughts are spatio-temporal patterns of coalitions of firing neurons and their interconnections. Neural algebras represent these patterns as formal algebraic objects, and a suitable composition operation reflects their interaction. Thus, a neural algebra is associated with any neural net. The present paper presents this formalization and develops the basic algebraic tools for formulating and solving the problem of finding the neural correlates of concepts such as reflection, association, coordination, etc. The main application is to the notion of consciousness, whose structural and functional basis is made explicit as the emergence of a set of solutions to a fixpoint equation.

Keywords: Neural nets, combinatory algebra, functional structures, emergent properties, models of consciousness.

1 Introduction

Thoughts are patterns of firing neurons. And so are sensory perceptions, memories, feelings, motor activations, etc. As a population of neurons fires in a pattern it causes the firing of neurons in another pattern, by its neural connections. Thus goes a broadly accepted view of the ongoing activity of the brain. The challenge met by this paper is to find a mathematical framework in which firing patterns M, N, etc. are the basic elements and their composition $M \cdot N = R$ describes how a pattern N which fires in a context M results in a pattern R. Indispensable for such an approach is that the mathematical objects M, etc. have a transparent relation to the basic neurological facts behind a spatio-temporal pattern of firing neurons.

Roughly speaking, each firing pattern is considered as being spread out into parallel tracks of successively firing individual neurons. Each of these tracks is understood as being divided into an initial part and a final part: if the track is part of the spread-out context M and its initial part belongs to N, then its final part belongs to $R = M \cdot N$.

K. Horimoto et al. (Eds.): AB 2008, LNCS 5147, pp. 96–109, 2008.

The goal of this paper is to cast this rough sketch into a mathematical model and to create the rudiments of a mathematical discipline for it. The proposed neural algebras provide the needed mathematical framework. It will serve for treating the problem of relating the functionality of a neural net to its communication structure in a coherent algebraic fashion.

The examples presented in this paper, in particular our excursion into a theory of consciousness, are necessarily simplistic, but they should show the spirit of such applications: the basic mathematical properties of neural algebras are used to formulate functional-structural problems as equations, and then to solve them algebraically. In particular we can locate neural correlates to some concepts that arise as descriptions of brain functions belonging to a higher level of descriptive language.

2 Neural Algebra

Neural algebras represent sets of neuronal activities – patterns, sequential and spatial, of firing brain cells. These correspond to neural activities as they would show up, for example, in sequences of functional MRI images. The formalization is then used to mirror the composition of such sets – one pattern causing other patterns – as a formal operation on the corresponding elements of the neural algebra.

The fascinating development of neurology, (as impressively told by Eric Kandel, "In Search of Memory" [11]), has allowed mathematical scientists to try approximating the anatomical, physiological and biochemical findings by more or less realistic models, so called artificial neural nets. The early story of these approaches and their relation to artificial intelligence ("connectionism") is collected in Anderson and Rosenfeld [1] reaching up to about 1987. There are other attempts to discuss activities of neural populations mathematically; the most promising so far is probably the dynamical systems approach; others, based on quantum mechanics are somewhat less convincing (see, e.g., the discussion in Koch and Hepp [12]).

Our approach is based on the representation of the activities of neural subpopulations as formal mathematical objects. Deciding on the level of detail about the functioning of neurons and their interconnection determines the interpretation of these expressions. Primarily, a neural net is the directed graph of its synaptic connections. The weakest description level approximates the functionality by providing weights to the synapses and introduces a discrete time behavior of neurons as known from artificial neural nets, familiar for their intuitive appeal, computational richness and developed theory. Our present exposition remains at this level; higher levels of detail might include specific biochemical, and perhaps other mechanisms of communication and distinguish various types of neurons, and might include stochastic elements.

To fix notation, an *artificial neural net* is a weighted directed graph, i.e. a triple (A, L, w), where the set $A = \{a_1, a_2, \dots\}$, which gives the name to the whole net, is a set of elements called *neurons*, connected by directed edges

$L : L \subseteq A \times A$ to which rational weights are attached by the function $w : L \to \mathbf{Q}$, abbreviated $w_{i,j} = w(\langle a_i, a_j \rangle)$. In an active neural net each neuron a_i has at any time instant t an *excitation value* $f(a_i, t) \in \{0, 1\}$, $t \in I$, a subinterval of discrete time $I \subseteq \mathbf{Z}$. These values are interrelated by the structure of the neural net as follows: If $\langle a_1, a_q \rangle, \ldots, \langle a_n, a_q \rangle \in L$, then

$$f(a_q, t+1) = H(\Sigma_i w_{i,q} f(a_i, t)) \,,$$

where H is a $0-1$ valued function: $H(v) = 1$ if $v \geq 1, 0$ otherwise.

Thus, neurons interact by sending information about their excitation states at time t along "axons" via "synapses" to other neurons. The synapses weigh these inputs and the receiving neuron derives from these inputs its excitation state at time $t+1$ according to the function H. At any given moment the weights $w_{i,j}$ attached to the synapse from neuron a_i to neuron a_j are given constants; "learning" may change these values, but this is outside the concern of this paper, although obviously important.

Models somewhat closer to physiological facts than this rather rudimentary one employ real-valued excitation functions and sigmoid functions H, see, e.g., Dehaene et al. [6]. Such models can be incorporated in our approach.

The basic building blocks of our theory are called *track expressions*, denoted by lower case symbols x, y, etc. These formal expressions denote the activation of specific neural connections at specific time instants as follows. For a single neuron a the expression consists of the symbol a alone. If neurons a_1 to a_k have directed edges to neuron a_0 and there is a further edge from a_0 to a_{k+1} and t is any time instant $t \in \mathbf{Z}$ then

$$x := \{a_1, \ldots, a_k\} \xrightarrow[a_0]{t} a_{k+1} \,.$$

is a track expression if the sum of weights of the incoming edges is at least 1.

The set on the left side may be empty, e.g. in the case that a_0 is an "input" neuron with no incoming edges. The neuron a_0 in a sense encodes the activation of this particular connection, it is therefore called the *key neuron* of this expression; formally $a_0 = \nu(x)$. The upper index t of the arrow indicates the time at which the key neuron is activated; formally $t = \tau(x)$. Any one of the a_i, say a_1, may itself be the key neuron of another connection, say

$$y := \{b_1, \ldots, b_s\} \xrightarrow[a_1]{t-1} b_{s+1} \,.$$

Then

$$\left\{ \{b_1, \ldots, b_s\} \xrightarrow[a_1]{t-1} b_{s+1}, \ldots, a_k \right\} \xrightarrow[a_0]{t} a_{k+1}$$

is also a track expression, still with a_0 as its key neuron. More track expressions are obtained by continuing the method of substitution as in the formal definition below.

Let A be any weighted directed graph. We formally define the set of track expressions x together with their key neurons $\nu(x)$ and firing time $\tau(x)$ as a set $S(A)$ of formal elements as follows:

$$
\begin{aligned}
S_0(A) \quad &= A,\ \nu(x) = \ \ x,\ \ \tau(x) = t \text{ for } x \in A, t \in \mathbf{Z} \\
S_{n+1}(A) = S_n(A) \ \cup \ \{ \ &\{x_1, \ldots, x_k\} \xrightarrow[a_0]{t} x_{k+1} : \textit{if there is an element} \\
&a_0 \in A \textit{ with edges from } a_1, \ldots, a_p \textit{ and to } a_q\,, \textit{such that} \\
&\tau(x_1), \ldots, \tau(x_k) = t-1, \tau(x_{k+1}) = t+1\,, \\
&\Sigma_i w(\nu(x_i), a_0) \geq 1\,, \\
&\{\nu(x_1), \ldots, \nu(x_k)\} \subseteq \{a_1, \ldots, a_p\} \textit{ and} \\
&\nu(x_{k+1}) = a_q, x_i \in S_n(A),\ i = 1, \ldots, k+1 \,\}\,, \\
\nu(\{x_1, \ldots, x_k\} &\xrightarrow[a_0]{t} x_{k+1}) = a_0\,, \\
\tau(\{x_1, \ldots, x_k\} &\xrightarrow[a_0]{t} x_{k+1}) = t\,.
\end{aligned}
$$

Then $S(A)$, the set of track expressions, is the union of the $S_n(A)$:

$$S(A) = \bigcup_{n=0}^{\infty} S_n(A)\,.$$

Iterated bracketing of track expressions serves to denote neural activities on increasingly higher levels of dependency. As theoretical constructs they are introduced to capture the compositionality of firing patterns and to thus facilitate the construction of an algebraic superstructure on a given neural net, the neural algebras.

We now come to the formal definition of firing patterns. Let A be a neural net, considered as a directed graph (equipped with further data in case of higher level detail as above), and let $S(A)$ be the set of all track expressions. A set M of track expressions constitutes a *firing pattern*, if, loosely speaking, it corresponds to a temporal pattern of firing neurons in A. Formally this means that there is an assignment of an excitation function $f(a_i, t)$ to the set of neurons a_i and firing times t occurring in the track expressions in M such that $f(\nu(x), \tau(x)) = 1$ for all track expressions x in M, and such that this assignment conforms to the firing laws.

Firing patterns are designed to be identified with (mental) functions as their *neural correlates*. This rests on the fact that certain subnets can be understood as having specific functionalities based on them. This singling out of subnets and firing patterns based on them is a virtual, theoretical, superstructure on the neural net and is typically guided by hypotheses on their function such as receiving sensory input or analysing activities based on some other subnet, etc. Research in neurology has resulted in an enormous and growing knowledge base of such facts for humans and for some animals.

Our mathematical framework identifies neural correlates as firing patterns; this makes it possible to capture the *compositionality* of such neural correlates

quite generally, as follows: Let the neural firing pattern M be based on a subgraph A_M of A, the *support* of M. Then M may take account of activities N supported by subgraph A_N and produce activities R, supported by A_R.

The activation of N allows the activation of key neurons in M, which in turn results, due to the structure of M, in the activation of the firing pattern designated by R. Mathematically this situation corresponds to a composition operation $M \cdot N = R$. Formally, we have the following definition:

$$M \cdot N = \{\, x : \textit{there is an element } \{x_1, \ldots, x_k\} \xrightarrow[a]{t} x \textit{ in } M \textit{ such that } \{x_1, \ldots, x_k\} \subseteq N \,\}.$$

Note that whenever M and N are firing patterns, then so is $M \cdot N$. This definition captures the rôles that the neurons in A_M and A_N play: Indeed, the neuron a has an activation history that depends on the histories of $\nu(x_1), \ldots, \nu(x_k)$ and influences that of $\nu(x)$.

It is in the nature of the things, that R itself may again be an "initiation" or an "action", etc. Indeed, each firing pattern can be used as a left multiplier, representing a law of interaction, or as a right multiplier, representing the input to the interaction. In this way, the set of firing patterns associated to a neural net A constitute an algebraic structure, the *Neural Algebra* $\mathcal{N}_A$.

3 Some Mathematical Background

Let A be a neural net. Let $F(A)$ be the set of firing patterns of A; this set is provided with the composition operation defined in the previous section, thus constituting an algebraic structure $\mathcal{N}_A = \langle F(A), \cdot \rangle$. It is in these algebras that we are to solve the equations describing interactions between neural processes formulated as firing patterns. Each one of these may of course contain neural populations that are not used in the composition operation and may, if we imagine them in nature, be physically far removed except for the overlapping necessary for the composition.

There are three useful theorems about neural algebras:

Theorem 1 (Fixpoint Theorem)**.** *In $\mathcal{N}_A$ all* fixpoint equations *have a solution; the solutions form a lattice by inclusion.*

Theorem 2 (Embedding Theorem)**.** *$\mathcal{N}_A$ is a subalgebra of a* combinatory algebra*, indeed, it is a combinatory algebra for certain nets A.*

Theorem 3 (Representation Theorem)**.** *If A is a sufficiently rich directed graph and Φ is a binary relation over B, a subset of A, then Φ is representable in $\mathcal{N}_A$ using an embedding f defined by: a and b are in the relation Φ if and only if $f(a) \cdot f(b) = f(b)$, where f maps B into $S(A)$.*

To prove Theorem 1, first note the monotonicity of the algebraic operation $M \cdot N$:

If $N_1 \supseteq N_2$ then $M \cdot N_1 \supseteq M \cdot N_2$ by the definition of the operation; equally $M_1 \cdot N \supseteq M_2 \cdot N$ for $M_1 \supseteq M_2$. Hence, if $\varphi(X)$ is any algebraic composition of

X with elements of $F(A)$ then $X' \supseteq X$ implies $\varphi(X') \supseteq \varphi(X)$. More generally, if D is a directed set of elements of $\mathcal{N}_A$ then $\varphi(\bigcup D) = \bigcup_{X \in D} \varphi(X)$. From this follows, that the fixpoint equation $\varphi(X) = X$ has a least solution $\bigcup_n \varphi^n(\emptyset)$, where $\varphi^0(X) = X$ and $\varphi^{n+1}(X) = \varphi(\varphi^n(X))$.

Namely: $\emptyset \subseteq \varphi(\emptyset) \subseteq \varphi(\varphi(\emptyset)) \subseteq \ldots$ is a directed set, hence $\varphi(\bigcup_n \varphi^n(\emptyset)) = \bigcup_n \varphi^{n+1}(\emptyset) = \bigcup_n \varphi^n(\emptyset)$; thus $\bigcup_n \varphi^n(\emptyset)$ is a fixpoint of $\varphi(X)$. In fact, it is the smallest fixpoint.

The above solution method for fixpoint equations can be expanded to simultaneous equations, e.g. in the case of the *coordination problem* (finding inputs that coordinate two activities M and N):

$$M \cdot P = Q\,, \qquad\qquad N \cdot Q = P\,.$$

Given M and N as known, the simultaneous fixpoint equations can be solved in $\mathcal{N}_A$ by a generalization of the method for one fixpoint: Let $P_1 = N \cdot \emptyset$, $Q_1 = M \cdot \emptyset$, $P_{n+1} = N \cdot Q_n$, $Q_{n+1} = M \cdot P_n$. Then $P = \bigcup_i P_i$, $Q = \bigcup_i Q_i$ are (least) solutions.

To simplify exposition, we occasionally drop the firing-time superscripts. We also may use the same track variables at different ocurrencies in an expression, the superscripts are thought of being supplied conforming to the actual positions.

To prove Theorem 2, let A be a complete directed graph: each node is connected to all nodes and all edges are weighted 1. Then the neural algebra $\mathcal{N}_A$ is a combinatory algebra, which means that it has the property that for any algebraic expression $\varphi(X_1, \ldots, X_k)$ in variables $X_1, \ldots, X_k$ there exists an element T in $\mathcal{N}_A$ for which $(\ldots((TM_1)M_2)\ldots M_k) = \varphi(M_1, \ldots, M_k)$ for all values $M_1, \ldots, M_k$ of $X_1, \ldots, X_k$. The object T is defined by

$$T := \{(\alpha_1 \to_{a_1} (\alpha_2 \to_{a_2} \ldots (\alpha_k \to_{a_k} x))) : x \in \varphi(\alpha_1, \ldots, \alpha_k)\,,$$

$$\alpha_1, \ldots, \alpha_k \subseteq S(A)\,,\ \textit{finite}\}\,.$$

It is traditionally called the *combinator* associated to the algebraic expression φ. To verify the case $k = 2$, consider $(TX_1)X_2 = \varphi(X_1, X_2)$:

$$\begin{aligned}(TM_1)M_2 &= \{x : \exists\alpha \subseteq M_1\,, \exists\beta \subseteq M_2\,, (\alpha \to_a (\beta \to_b x)) \in T\} \\ &= \{x : \exists\alpha \subseteq M_1\,, \exists\beta \subseteq M_2\,, x \in \varphi(\alpha, \beta)\}\,.\end{aligned}$$

The last equation follows from

$$\varphi(M_1, M_2) \;=\; \bigcup\{\varphi(\alpha, \beta) : \alpha \subseteq M_1\,,\ \beta \subseteq M_2\,,\ \alpha, \beta \text{ finite}\}\,.$$

proving Theorem 2.

The proof of Theorem 3 consists of verifying the following set-recursive definition of the mapping f, for all a and b in the relation Φ:

$$f(a) = \{a\} \cup \{\{b\} \longrightarrow_a x : b \in B, x \in f(a)\}\,.$$

Neural algebras, as we have defined them, are related to combinatory algebras, as shown above. The latter have evolved from beginnings in mathematical logic,

namely Lambda Calculus and the related Combinatory Logic [2,4]. While these subjects were created in the 1930's with the foundations of mathematics as their aim, they have had considerable influence in theoretical computer science, especially after Scott and Plotkin constructed their well-known models of the Lambda Calculus. (For a concise introduction see Engeler [10], Chapter 3.) The basis of the present work is a richer type of models, the subject of a prolonged effort of the author and his students at the ETH: "The Combinatory Programme", [9]. This research program deals with a large variety of mathematical subjects, including universal algebra and computer algebra, and later set theory and category theory.

Neural algebras as they arise from natural examples do not have complete graphs, although they have very large numbers of synaptic connections indeed. In applications, therefore, the neurons needed for realizing firing patterns like the T above in the proof of Theorem 2 may have to be obtained by enlarging the underlying graph A to B; "recruiting new neurons and synapses" as we may say. Mathematically speaking, this corresponds to an algebraic extension $\mathcal{N}_B$ of the original $\mathcal{N}_A$ which then contains the new element. This is a construction familiar in algebra, where to solve equations it may be necessary to expand the algebraic structure, e.g. from rational to algebraic numbers. In natural neural nets, such expansions may conceivably consist in mobilizing already present but partially dormant neurons and connections.

The case of *concordance* of activities is an example: if U and V are two firing patterns, their intersection $U \cap V$ describes their functional and temporal concordance, the extent to which U and V concur. $U \cap V$ is in fact the result of applying the operator $\wedge$ on them:

$$\wedge UV = (\wedge \cdot U) \cdot V = U \cap V \,,$$

with

$$\wedge = \{\{x\} \longrightarrow_r \{\{x\} \longrightarrow_s x\} : x \in S(B)\} \,,$$

where r and s are two newly recruited neurons. The union of two firing patterns can be obtained in a similar manner.

The solution of equations other than fixpoint equations is a challenging mathematical problem. Indeed, it can be shown that all degrees of computational complexity and of unsolvability can occur. Some solution methods based on algebraic extensions of a combinatory algebra have been described in Chapter III of [9], but much work needs to be done and experience gathered from applications.

4 In Search of Consciousness

One possible application of our theory is in the search for neural correlates of mental functions. Let us turn to the entirely speculative case of "consciousness", with the goal of analyzing the well-known thesis that consciousness is the power of self-reflection. The definition of consciousness as self-reflection is just one of a long and involved history of attempts to define this concept. In this, we are well aware of the caveat of Francis Crick: "Until the problem [of consciousness]

is understood much better, any attempt at a formal definition is likely to be either misleading or overly restrictive, or both," [3]. For one thing, the sheer size and complexity of that network precludes any complete analysis. More fundamentally: There can be no effective method to decide whether a neural net, once initiated, will develop a given response. (This can easily be shown by simulating Turing Machines in neural nets).

However, it is quite possible to work out attributions of content and function to types of neural structures; in the case of consciousness this results in the *Structure Theorem* below.

Let us then understand neural consciousness as *the ability of a neural net B ("the brain") to consciously observe itself as being conscious and as consciously planning and acting.* These abilities are embodied as activities in sub-populations of the "brain", to be represented here by firing patterns; their interrelation is expressed by their composition: If C is the firing pattern corresponding to "consciousness", and M_1, M_2, etc. are the firing patterns corresponding to the context of observing, acting, planning, moving, etc. then $M_1 \cdot C$, $M_2 \cdot C$, etc. are the results of observing, acting, etc. as dependent on consciousness. To the sum of these results, together with C itself, C is again applied. Translated into neural algebra, our definition of *consciousness* transforms into an equation of the form

$$C \cdot \left(C \cup \bigcup_i M_i \cdot C \right) = C \,.$$

The solutions of this fixpoint equation constitute the set of persistent activity patterns in a net of neurons that may be understood as states of "consciousness". (The apparent circularity of our non-formal definition thus resolves itself as multiple entry of the unknown in a single equation).

Obviously, the *quality of consciousness* in this formal sense depends crucially on the size and structure of the underlying "brain" and on the degree of involvement of other brain functions, such as memory, language, intentionality. This is reflected in the fact that different such activity patterns are (lattice-)ordered by inclusion and correspond to to different forms or to emerging stages of consciousness. Thus, forms of consciousness may already be found in neural nets of primitive animals, and indeed even in neural simulations of computers.

Basically, our results say that consciousness is always based on one or more recurrent loops of active neurons and feeds forward from these to other activated regions of the brain; patterns that are solely based on stimulus-and-response cannot support consciousness. The various forms of consciousness depend in this way on the richness and the activity of the mind as embodied in the neural net constituting the brain.

4.1 The Structural Basis of Consciousness

To obtain the structural basis of consciousness, we solve the fixpoint equation *structurally*, which means that we "disregard time". Formally, this means that

$$\{x \underset{a_1}{\longrightarrow} y : x \in A_3,\, y \in A_2\}$$

is to be read as

$$\{x \xrightarrow[a_1]{t} y : x \in A_3,\, y \in A_2,\, t \in \mathbf{Z}\}\,.$$

Computing consciousness fixpoints shows up the structural facts that are relevant in all models of the brain. We observe the following facts:

Theorem 4 (Structure Theorem of Consciousness)
(a) Consciousness always has a base in one or more cycles of the directed graph.
(b) Consciousness can be expanded along any outgoing edge.
(c) Consciousness never expands backwards into cycle free "stimulus and response" subgraphs.

To prove (a) consider the simple case of a cycle of neurons a_1, a_2, a_3, cyclically connected with weights 1. Assume this cycle embedded in a graph with an edge of weight 1 leading from a neuron b to a_1, and one from a_3 to a neuron c, again with weight 1.

Let

$$A_1 = \{a_1\} \cup \{x \underset{a_1}{\longrightarrow} y : x \in A_3, y \in A_2\}\,,$$
$$A_2 = \{a_2\} \cup \{x \underset{a_2}{\longrightarrow} y : x \in A_1, y \in A_3\}\,,$$
$$A_3 = \{a_3\} \cup \{x \underset{a_3}{\longrightarrow} y : x \in A_2, y \in A_1\}\,,$$
$$C = A_1 \cup A_2 \cup A_3\,,$$
$$B = C \cup \{b \underset{a_1}{\longrightarrow} y : y \in A_2\} \cup \{x \underset{a_3}{\longrightarrow} c : x \in A_2\}\,.$$

Then

$$C \cdot C = A_1 \cdot C \cup A_2 \cdot C \cup A_3 \cdot C = C\,,$$

and

$$B \cdot C = C \cdot C \cup \emptyset \cup \{c\}\,.$$

Hence

$$C \cdot (C \cup B \cdot C) = C \cdot (C \cup \{c\}) = C \cdot C = C\,.$$

Items (b) and (c) can be as easily proved: (b) is exemplified by extending C to $C' = C \cup \{x \underset{a_3}{\longrightarrow} c : x \in A_2\}$; (c) by observing that $C = C \cup \{b \underset{a_1}{\longrightarrow} y : y \in A_2\}$ is not a fixpoint.

4.2 The Emergence of Consciousness

Consciousness "simply happens" in any sufficiently rich neural net. It is a typical example of an *emerging phenomenon*. Mathematically, emergence consists in the approximation of a fixpoint: Assume that we already have a fixpoint C (the empty set $\emptyset$ is always available), and let E extend C and N extend M. Using E and N we can gradually progress to a new fixpoint C' as follows:

Define $\varphi_Y(X) = X \cdot (X \cup Y \cdot X)$ and observe

$\varphi_M(C) = C \subseteq E \subseteq \varphi_N(E) \subseteq \varphi_N(\varphi_N(E)) \subseteq \ldots$ is a directed set, hence $\varphi_N(\bigcup_n \varphi_N^n(E)) = \bigcup_n \varphi_N^{n+1}(E) = \bigcup_n \varphi_N^n(E)$; thus $C' = \bigcup_n \varphi_N^n(E)$ is a fixpoint of $\varphi_N(X)$.

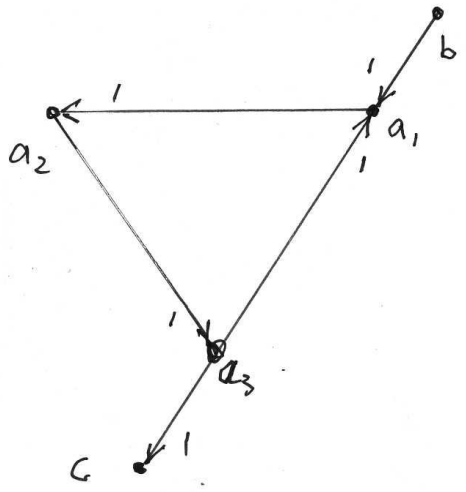

Fig. 1. A cycle of consciousness

For a somewhat plausible example, imagine a chess player whose sensory system provides him with the positions in the game, upon which his planning faculty decides on a plan, e.g. a particular endgame. The action coordination determines the first move, which the motorics of the player transforms into moving a particular piece. There follow updates of the planner about that move and to remember to wait for the challenge for his next move.

Using this context as a guide, consider a *microbrain* $\mathcal{N}_B$, the neural algebra over the directed graph B, representing the neural substrate which supports firing patterns which we name, to fix ideas, I for sensory input, V for vision, P for perception and planning, A for activation of actions, M for motor activity, L for language, S for speech, H for body perception. In our microbrain, these firing patterns have as key neurons just one neuron each, namely i, v, p, a, m, l, s, h, connected and weighted according to the diagram in Figure 2. The structural laws for the firing patterns of $\mathcal{N}_B$ can be read off the diagram in Figure 2 as follows in the form of simultaneous recurrence equations:

$$H = \{h\} \cup \{x \longrightarrow_h y, x \longrightarrow_h z : x, y \in H, z \in A\}$$
$$A = \{a\} \cup \{x \longrightarrow_a y, x \longrightarrow_a z, u \longrightarrow_a z, \{x, u\} \longrightarrow_a z_1, \{x, u\} \longrightarrow_a z_2$$
$$: u \in H, x, y \in P, z_1 \in M, z_2 \in L\}$$
$$M = \{m\} \cup \{x \longrightarrow_m y : x \in A, y \in V\}$$
$$V = \{v\} \cup \{x \longrightarrow_v y, z \longrightarrow_v y, \{x, z\} \longrightarrow_v y : x \in M, y \in P, z \in I\}$$
$$P = \{p\} \cup \{x \longrightarrow_p y, z \longrightarrow_p y, z \longrightarrow_p u, x \longrightarrow_p u : x, y \in A, z \in V, u \in L\}$$
$$L = \{l\} \cup \{x \longrightarrow_l z, y \longrightarrow_l z, \{x, y\} \longrightarrow_l z : x \in P, y \in A, z \in S\}$$
$$I = \{i\}, S = \{s\}$$

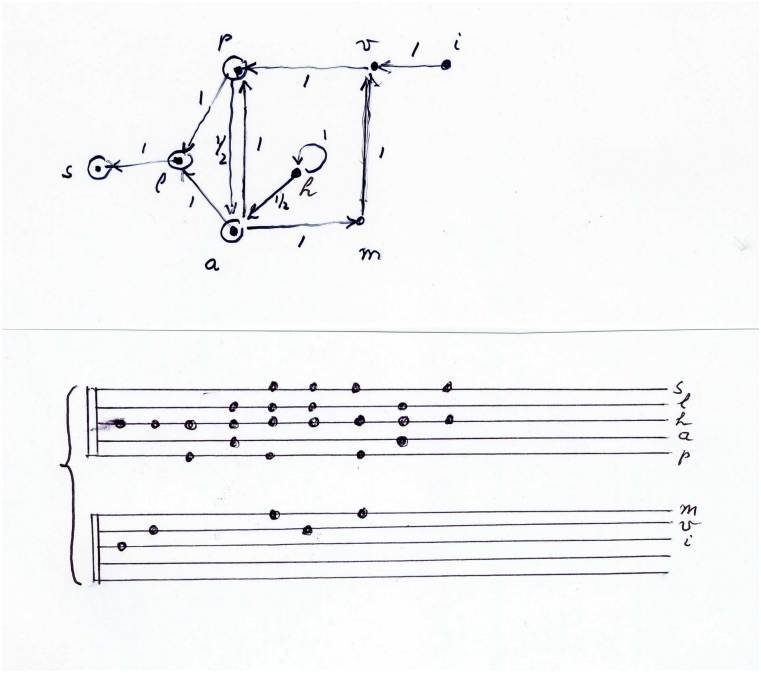

Fig. 2. The microbrain $\mathcal{N}_B$ and its scores

The consciousness equation will admit at least one but conceivably many solutions in $\mathcal{N}_B$. Expanding them, as suggested by *emergence*, may add further ones.

The minimal solution is, of course, the empty set $\emptyset$, since $\emptyset \cdot X$ is the empty set for all X.

A first nontrivial solution is $C_0 := H_0$ "the brain may only be conscious of the beating heart". Let, again recursively,

$$H_0 = \{h\} \cup \{h : x \longrightarrow_h y : x, y \in H_0\}\,.$$

To verify that H_0 supported by the cyclic subgraph on the neuron h is indeed a fixpoint follows at once from (a). Other fixpoints follow as consequences of (b) and (c).

As observed above, consciousness-fixpoints in $\mathcal{N}_B$ form a lattice under the inclusion relation and are as a rule not all ordered in one sequence. They are best described by the supporting subgraphs of these patterns, e.g.

$$C_1 = A_1 \cup P_1 \cup H_1\,,$$

based on key neurons h, a and p, where

$$A_1 = \{a\} \cup \{\{x, y\} \longrightarrow_a z : y, z \in P_1, x \in H_1\}\,,$$
$$P_1 = \{p\} \cup \{x \longrightarrow_p y : x, y \in A_1\}\,,$$
$$H_1 = \{h\} \cup \{x \longrightarrow_h y, x \longrightarrow_h z : x, y \in H, z \in A_1\}\,.$$

Remaining with microbrains, we may also incorporate other ideas about the functioning of the brain. For example, we may interpolate a neuron w between p and v, which makes the system "watch out" for particular input. Or we may establish an edge between v and m for immediate reactive movements, etc.

Even in rudimentary neural algebras such as $\mathcal{N}_B$, there develops a rich variety of conscious behavior. The activation history of such a net may be looked at as a sort of musical score for the "theme" that the brain plays, see Figure 2 for an example. The activation of neurons on which consciousness is based is shown at the top of the score, the bottom would be something like (relative) subconscious.

Visualization of consciousness as "a sort of orchestral piece played in the brain" points out a connection with the findings of the school of Wolf Singer (e.g. [8]), which shows the central importance of synchronicity. It also shows the importance of the persistence of activation patterns for the constitution of a sustained consciousness, the "self" in a "*sin*-phonic" view, possibly with a recognizable personal "style".

5 Discussion

To belabor the obvious: as a model of the human brain, the microbrain $\mathcal{N}_B$ is unrealistic by about twelve orders of magnitude; the neuron v for example stands for something like the visual cortex, and m may be a cascade of interrelated neural nets, etc.

However, our findings about the existence of the lattice of consciousness activity complexes scale up to "brains" of all levels of complexity, and we may speculate, whether "core consciousness" and levels of "extended consciousness", as described in the literature (e.g. by [5]), correspond to such fixpoints. Also, we may speculate about the history of activations of these different forms of consciousness and whether this might involve moving up from body-awareness such as C_1 based on lower fixpoints to higher fixpoints. Experiences by introspection might just be based on such migrations, which, by the way, may have to move through higher or lower points in the lattice to reach from one to the other. Other such experiences (e.g. so-called "earworms" of popular melodies), as well as observations on periodicities of brain activities in observed conscious behaviors, seem to match the musical-score paradigm of brain activity.

In neural algebra, thoughts, emotions, communication, etc., are just elements to be computed with, this is all there is, *formally*. Remaining with the musical score paradigm, the notes of a Mozart piece would analogously be all there is! But, far from formality being an impoverishment of these concepts, the mathematical approach presents an unending challenge, of which we now sketch only a few immediate aspects.

5.1 On Laws of Thought

When Boole created an algebraic discipline for computing with truth values of statements, "thoughts" were understood as being expressed in language, and

parts of the grammar of language provided the patterns of the algebraic operations. Should we not now take firing patterns as "thoughts" and their composition as the algebraic operations? This results, abstracting from the unknown complexities of human neural algebras, in a position which we could call *neural logic*, regarding neural algebra as the true algebra of thoughts.

If we aim to understand mental activities as compositions of firing patterns, there are several basic concepts that need neural correlates. Further research would have to show, whether, for example, the following proposals have a chance:

(a) To classify a firing pattern X as conforming to a template F, as in recognizing a face, we could simply compose the two and consider the result as indicating in what, and how far, the classification holds true. Or, taking the basic idea from the Representation Theorem 3, we might identify true classification with $F \cdot X = X$.
(b) In Theorem 2 it is shown that complex composition patterns of objects can be considered as objects in a combinatory algebra. In the present context, we might consider the notion of analogy as a particular firing pattern, say L. Then L expresses the fact that the notion X is to the leading example U as it is to the analogon V, thus:

$$LXUV = \wedge \cdot (X \cdot U) \cdot (X \cdot V)\,,$$

using the intersection operator from Section 3.
(c) Combinatory algebra has objects that correspond to natural numbers and to computing with them; our model of the brain inherits this. Although this shows that a rich enough $\mathcal{N}_B$ can handle all computable functions, and indeed simulate any Turing machine, it is implausible that the published versions of arithmetic in combinatory algebra (as in [10]) are the ones realized in the human brain, (cf., for example, Dehaene [7] on numerosity). More generally, "intelligence" may be closely related to the easy availability of combinatory objects, templates, which represent basic forms of relatedness, such as analogy, causality, duality, etc.

5.2 Extended and Collective Consciousness

In Section 4 we have taken the simplest case of composition of "consciousness" with the structure of the mind, namely

$$C \cdot \left(C \cup \bigcup_i M_i \cdot C \right) = C\,.$$

The "mind" $M = \bigcup_i M_i$ may allow artificial extensions $M' \supset M$, which may conceivably induce an extension of consciousness. The use of tools "till they become a part of ourselves" is a striking example. In designing tools, e.g. software tools, it may therefore be of interest to keep in mind the neural connections that we have identified with the structural functioning of consciousness.

There is no reason to restrict the present approach to neural nets that are "brains". Equally well we could consider populations of agents, for example a colony of ants, that operate under certain constraints and with certain well-defined schemes of communication. Thus would emerge concepts of collective consciousness for such populations; and it would be attractive to try and develop this idea in a variety of contexts. In a similar vain, the human brain itself is in fact a population of such individual agents, neurons, whose collaboration may have evolved by a learning process, including the recruitment of new members and new connections.

Acknowledgments. I wish to express my thanks to Klaus Hepp, who challenged me on my bold statements that I knew "the right mathematics for modelling the brain and consciousness." In consequence, the present approach was worked out and presented in a preliminary form at the "Brain Fair Zurich" in 2005. Of course, all responsibilities for this work remain mine.

References

1. Anderson, J.A., Rosenfeld, E.: Neurocomputing. MIT Press, Cambridge (1988)
2. Church, A.: A Set of Postulates for the Foundations of Logic. Annals of Math. 33, 346–366 (1932)
3. Crick, F.: The Astonishing Hypothesis. Simon & Schuster, Ltd (1994)
4. Curry, H.B.: Grundlagen der Kombinatorischen Logik. Amer. J. of Math. 52, 509–536, 789–834 (1929)
5. Parvizi, J., Damasio, A.: Consciousness and the Brainstem. Cognition 79, 135–159 (2001)
6. Dehaene, St., Sergent, C., Changeux, J.-P.: A Neural Network Model Linking Subjective Reports and Objective Physiological Date During Conscious Perception. Proc. Nat. Acad. Sci. 100, 8520–8525 (2003)
7. Dehaene, St., Molko, N., Cohen, L., Wilson, A.J.: Arithmetic and the Brain. Current Opinion in Neurobiology 14, 218–224 (2004)
8. Engel, A.K., Fries, P., Singer, W.: Dynamic Predictions and Synchronicity. Nature Reviews Neuroscience 2, 704–716 (2001)
9. Engeler, E., Aberer, K., Gloor, O., von Mohrenschildt, M., Otth, D., Schwärzler, T., Weibel, T.: The Combinatory Programme. Birkhäuser, Boston (1995), `http://www.math.ethz.ch/~engeler`
10. Engeler, E.: Foundations of Mathematics, ch.3. Springer, New York (1983) also in Russian, Chinese and German
11. Kandel, E.R.: In Search of Memory, Norton New York (2006)
12. Koch, C., Hepp, K.: Quantum Mechanics in the Brain. Nature 40, 1011–1012 (2006)