TICAM REPORT 97-11
June, 1997

Consciousness is an Information Field Induced by Hebbian Dynamics

Willard L. Miranker

Consciousness is an Information Field [1]

Induced by Hebbian Dynamics

by

Willard L. Miranker[2]
Department of Computer Science
Yale University

To Terri Sisson, muse, on the occasion of the thirtieth of May.

Abstract: We introduce a field of information associated with the action potentials, the latter encoding conventional unconscious neural processing. We show that the field generates a dual representation of the neural processing which mirrors the information states conveyed by the neural circuitry itself. This leads to the claim that the field is the conscious experience of neural processing. An explanation for the fitness advantage of consciousness in evolution comes as a by-product. We start with the Hebbian synapse whose dynamics are interpreted as an atom of awareness, and which is quantified in terms of the signature of the time rate of change of synaptic strength. We show how the information field is built up out of such atomic constituents. The mathematical development shows that consciousness (i.e., the field) arises through a coupling of Russelian-like internal (the dual information field) and external (the primal action potential) properties of matter. This is contrasted with a corresponding duality in quantum mechanics where consciousness itself enters as a causal agent. The model presented is falsifiable, and a method for verifying it experimentally is suggested.

[1]This manuscript was written during a stay at TICAM, the University of Texas at Austin, March, 1997. The author is grateful to T. Oden and R. v.d.Gejn for helping to make this stay possible.

[2]Research Staff Member Emeritus, IBM T.J. Watson Research Center, Yorktown Hts., N.Y.

1 Introduction

We propose that *conscious experience* corresponds to a *field of information* which accompanies neural processing. The field is associated with, but is different from, the action potentials in terms of which conventional neural processing itself is conducted. The information field varies in strength, depending upon details of the neural processing, and when it approaches its (normalized) maximum value, the information field is the conscious experience of the neural processing which it parallels. Thus our approach is a coupling of internal (the information field) and external (the neural processing) properties of matter (B. Russel, 1927). There is a duality in quantum mechanics in which consciousness itself enters as a bridge between primal and dual, namely as a causal agent in the so-called collapse of the wave function from which a measurement emerges. By contrast, in the presentation here, the causal agent is a physical threshold effect, and it is consciousness which emerges.

We start by reviewing the standard Hebbian synapse, and we interpret its dynamics as an *atom or quantum of awareness.* This atomic awareness is expressed as information which is measured in terms of the signature of $\dot{s}$, where s is the synaptic strength. The field of information is supported by a collection of neurons, and the field's value at any one of those neurons is a function of the information contained in that neuron's set of afferent synapses. We take the field value to be the average over this set. We show that this average, that is, the information field's value is correlated to the action potential itself and varies in strength according to the degree of correlation within the neuron's set of afferent synaptic activities. Thus we shall see that in some circumstances, the hypothesized field is an exact correspondent of the unconscious signals being conveyed and processed by a collection of neurons in the customary sense.

Since the field stems from an atomic awareness (in the Hebbian synapse), since it mirrors the unconscious information processed by collections of neurons, and since it increases in strength as the degree of correlation among the neural inputs increases, we shall hypothesize that this new field is consciousness itself. A by-product of our approach is an explanation of the fitness advantage of consciousness in evolution.

The ideas presented here were motivated in part by P. Hut and R. Shepard, 1996. They speculate that to explain consciousness a new property 'X' which stands to consciousness as time stands to motion is needed. Here we formulate such a property, namely information. The new dimension has aspects which place it in between an independent variable (such as time) and a dependent variable (such as mass).

No theory of consciousness has been accessible to experimental validation. To validate the present hypothesis, it would be sufficient to measure consciousness through the construct which we develop. We conclude with a suggestion for doing this.

Taxonomy

Using the terminology of D. Chalmers, 1995, this presentation is concerned with the *hard problem of consciousness*, the problem of explaining experience. Theories of

consciousness have been taxonomized into *Materialist Theories* of types A and B (D. Chalmers, 1997), and *Mysterian Theories* (V. Hardcastle, 1996). Type A materialist theorists (for example, F. Crick and C. Koch, 1995 and P. S. Churchland, 1996) do not recognize the existence of the hard problem. They claim that the phenomenon of experience can be reduced to known physical laws. The type B materialists (for example, V. Hardcastle, 1996) recognize the existence of the hard problem but expect it to be explainable by known physical laws. (See W. Miranker, 1997 also.) The mysterians (for example, R. Penrose, 1989 & 1994, S. Hameroff and R. Penrose, 1996 and D. Chalmers, 1996) believe that new physical laws are needed to address the hard problem. It is possible to argue that the presentations made here belong to each of the three categories: type A materialism, type B materialism, and mysterianism.

In Section 2, we introduce the information field and develop some of its properties, including, in particular, a threshold property (gain) which we interpret as a basic emergence (of consciousness) effect. In Section 3 we comment on additional aspects such as (i) quale, (ii) the primal/dual nature of our model and the contrasting of this with a duality in quantum mechanics, (iii) degree of consciousness, (iv) binding, (v) an information theoretic property of consciousness with its Darwinian implication, (vi) machine consciousness, and finally (vii) we propose a method of measuring consciousness, demonstrating the falsifiability of the model presented here. Along with the last we include a word of taxonomic critique.

2 The Information Field

For reasons of clarity, we shall deal with a simple model, the McCulloch–Pitts neuron with n input synapses. Then let v^e be the efferent neural activity of such a neuron. Let $v^a = (v^a_1, \ldots, v^a_n)^T$ be the vector of afferent neural activities, and let $s = (s_1, \ldots, s_n)^T$ be the vector of corresponding synaptic strengths. We write the Hebbian synaptic dynamics for each neuron as the following system of n differential equations,

$$\dot{s} = H(v^e, v^a).$$

Here H is the so-called Hebb function. (See E. Kairiss and W. Miranker, 1997 for further details concerning Hebbian synaptic dynamics.)

As is customary for the McCulloch–Pitts neuron, we take the neural output to be

$$v^e = h(s \cdot v^a - \theta),$$

where

$$s \cdot v^a = \sum_{j=1}^{n} s_j v^a_j,$$

$$h(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

and $\theta > 0$ is a threshold.

The metaphor of experience

We interpret the Hebbian dynamics as an atomic awareness. The mediation of the value $H(v^e, v^a)$ is unknown, and we view it as metaphysical. That is, the Hebbian dynamics are taken as a postulated and unreducible property of nature (in category, analogous to the law of gravity, say).[1] We say that the synapse 'experiences' v^a and v^e, and the nature of the experience is to increase or decrease the value of s. That is, it is the signature of $\dot{s}$ which potentiates the experience (of the synapse). Further development of the ideas here could very well be concerned with the magnitude of $\dot{s}$.

The analog of this viewpoint in the context of gravity, say, is that one mass 'experiences' the presence of a second, and the nature of that experience is to change the distance between the masses according to Newton's third law. That is, it is the gravitational force which is 'experienced'. As far as we know, gravity and the third law are metaphysical. They are postulated and unreducible properties of matter. The gravitational analogy can be drawn even closer to the Hebbian by considering Keplerian motion (H. Corben and P. Stehle, 1950). Take the case when an elliptical orbit is executed by one of the masses with distance r to the other. Between the apses of this orbit the signature of $\dot{r}$ is invariant. We might say that the mass pair experiences the signature of $\dot{r}$. That is, the pair experiences the separate attractive and repulsive stages of the motion as separate sensations. Viewed in this context we see that our synaptic/information-field approach is a coupling of internal and external aspects of matter (B. Russel, 1927).

Specification of the field

We associate an information vector I with a neuron, the j-th component of I being the information associated with the j-th afferent synapse, as follows.

$$I_j = \frac{1 + \mathrm{sig}\dot{s}_j}{2}.$$

Note that v^e is a binary scalar, v^a and I are vectors of such scalars, s and $\dot{s}$ are real valued vectors.[2] We write I as the Boolean 'and' function of v^e and v^a, where componentwise

$$\begin{aligned} I_j &= v^a_j \wedge v^e \\ &= v^a_j \wedge h(s \cdot v^a - \theta). \end{aligned}$$

Corresponding values of v^e, $\dot{s}$ and components of v^a and I are summarized in the following table.

[1] Hebbian dynamics might in the future become expressible in terms of underlying processes. In that case we should apply the words metaphysical and unreducible in these last two sentences to those processes as appropriate, leaving this presentation otherwise essentially unchanged.

[2] Returning briefly to the Keplerian metaphor, we could identify the separate attractive and repulsive sensations there with the two values of information, $(1 + \mathrm{sig}\ \dot{r})/2 = 1$ or 0. Indeed this suggests an alternate to the terminology 'information', namely 'sensation'.

v_j^a	v^e	$\dot{s}$	I_j
1	1	>0	1
1	0	<0	0
0	1	<0	0
0	0	0	0

Let $\bar{I}$ denote the average of the components of I:

$$\bar{I} = \frac{1}{n}\sum_{j=1}^{n} I_j = \frac{1}{n}\sum_{j=1}^{n} \frac{1+\mathrm{sig}\dot{s}_j}{2}.$$

Recall that $10^3 \leq n \leq 10^5$ in the human brain. The binary vector v^a takes its values at the vertices of the unit n-cube. Let the number of nonzero components of v^a be denoted by

$$|v^a| = \sum_{j=1}^{n} v_j^a.$$

$|v^a|/n$ is the relative number of afferents which are firing. Equivalently $|v^a|/n$ is the specific input intensity.

We take the value of the information associated with the neuron to be $\bar{I}$. Since $v^e = 0$ or 1, we see that

$$\bar{I} = \frac{1}{n}\sum_{j=1}^{n} \left\{ \begin{matrix} 0 \\ 1 \end{matrix} \right\} \wedge v_j^a = \left\{ \begin{matrix} 0, & v^e = 0, \\ \frac{|v^a|}{n}, & v^e = 1, \end{matrix} \right.$$

with $\bar{I} \to v^e$ ($= 1$ or 0), increasing or decreasing, as the case may be, as the correlation of the afferent synaptic activity increases. That is, the information is zero if the neuron does not fire, and it is equal to the specific input intensity if the neuron does fire. Thus the information quantity $\bar{I}$ mirrors v^e. Details of the derivation of this expression for $\bar{I}$ in the special cases $n = 1$ and 2 are given in the Appendix.

Collections of neurons

Although we have focussed the derivation here on a single neuron, it is critical to note that neural processing, both conscious and unconscious, is the function of large collections of neurons. The information comprises a field supported by such large collections. What we have derived is the field's value $\bar{I}$ at one position, at one neuron.

Gain and emergence

Note that gain, namely the threshold effect for the generation of v^e is carried over to $\bar{I}$ by this definition. In Section 3 we shall interpret the field as consciousness, and so, we recognize this threshold behavior as the characteristic 'emergence quality' of consciousness.

3 Comments

We now give a number of interpretive comments, connecting the construct of the model with several expected properties of consciousness and with one novel information theoretic property. We conclude with a suggestion regarding experimental verification.

Quale

We have seen that the value of $\bar{I}$ for any McCulloch-Pitts neuron approximates the action potential of that neuron, the more closely, the greater the correlation among the neuron's afferent synaptic activities. Let us simply call 'scene', the physical information (an image, a sound, an odor, a pain, ...) which is at any instant of time being processed by a collection of neurons. This scene is encoded, is represented by the neural activity of the collection. We shall refer to this conventional representation of the scene as the *primal version*, and we stress that it is unconscious. The information field is an alternate, a *dual representation* of the scene.

What we do here is to hypothesize that this dual representation is consciousness. This being the case, we may use the term qualia for the dual version of a scene.

Duality and quantum mechanics

Let us contrast this consciousness/unconsciousness duality (i.e., this qualia/scene duality) with one of the dualities in quantum mechanics. Quantum mechanics consists of a physical-like part and an experiential part. The former consists of the waves of probability amplitude and the Shrödinger dynamics which propagates these waves. This physical part contains all of the *objective tendencies* (the *potentia* of Heisenberg) for transition from the possible (primal) to the actual (dual). The physical (primal) aspect of quantum mechanics stands in correspondence to the physical part of the theory here, namely to the conventional unconscious processing of signals in the neural circuitry.

The dual part of quantum mechanics is based on experiencing nature, that is on measurement, according to Bohr. More explicitly, according to J. von Neumann, 1955 and E. Wigner, 1961, it is consciousness itself which is the causal agent for the collapse of the wave function and the emergence of a classical result (a measurement). The dual part of our theory is the emergence of the information field from the 'potentia' of the conventional states in the neural circuitry. The causal agent for this is a high degree of correlation among the afferents in an entire collection of neurons. This spawns the action potentials of those neurons, each by means of a threshold effect, and it induces the appearance of the information field. For us the causal agent is a physical effect and consciousness itself emerges as the information field.

It is of interest to compare these observations with theories of consciousness based directly on quantum mechanics as in H. Stapp, 1996.

Degree of consciousness

There is a degree of consciousness built into this construct. Namely as the correlation (among the afferents) referred to weakens, the fidelity of the approximation of

the primal by the dual weakens (the latter being an average of the afferent activity). At some point, perhaps at only slight departure from perfect correlation of the afferents, the consciousness which is weakening disappears altogether. Compare this with the comment on gain which concludes Section 2.

Binding

This weakening/disappearing behavior suggests an explanation for why we can be conscious of very few things at once. The explanation needs to lie in the neural connectivity, in the wiring. Namely as one collection of neurons (corresponding to one experience of one scene, i.e., corresponding to one qualia) has all of its neurons' afferents become respectively, highly correlated, it might be that there is a corresponding inhibitory effect which disrupts (weakens) correlation in neighboring collections. (Actually one collection of neurons could support a dynamic repertoire of conscious experiences (of quale) with a winner-takes-all protocol. That is, one conscious experience, will subordinate all others in the repertoire by means of the disrupting inhibitory effects. In a sense the neural collection is tuned to one experience at a time by the correlation/inhibition.) This reflects on the so-called binding problem of consciousness. The correlation employed by this model could crystallize over several cortical regions and involve many thousands of neurons and many millions of synapses. We speculate that this provides a stage of sufficient capacity for binding the several hetero-sensory inputs required for a conscious experience.

Infomax and evolution

The correlation just referred to has an infomax interpretation. That is, the greater the quantity $|v^a|$ (i.e., the greater the redundancy among the components of v^a), the greater is the mutual information in the network driving the neuron in question. (Recall that mutual information $I(v, x)$ of an input/output system subject to noise is the reduction in uncertainty about the input x given the output v. While this mutual information $I(v, x)$ is associated with the information field, the two are different quantities.) Indeed, consider a collection of N neurons, each with the same input $x = (x_1, \ldots, x_n)^T$. Let us focus for the moment on $v^a = (v_1^a, \ldots, v_N^a)^T$ as a vector of different neuronal outputs. Here v_j^a is the output of a neuron j, $j = 1, \ldots, N$. The average mutual information $I(v_j^a, x)$ for each neuron in this collection is increased with redundancy in the components of the vector v^a (under a set of conditions on the gain, on the type and independence of the input noise, etc. See S. Haykin, 1994, pp. 455-8 for details. See R. Linsker, 1988 also.) Of course, it is the increasing redundancy in the components of v^a which leads to the emergence of consciousness, according to the arguments here.

Thus according to our interpretation, *consciousness is a mechanism associated with increasing the mutual information in a network.* This suggests the fitness advantage of consciousness in evolution.

Can machines be conscious?

Information processing machines are by construction supplied with a primal system of representation of scenes. To date, scenes in a digital computer correspond

to arrays of bits. Analog devices, such as artificial neural networks may have more interesting scenes. To create consciousness in a machine, according to the approach taken here, we must arrange to begin with that its processing elements have the capacity to induce or generate a dual system of information representation. This seems unlikely for the digital computer as it is currently constituted. Artificial neural nets have plastic synapses, and so it may be possible to build them with the property needed to support the effects which we have described.

We are a long way from being able to build an artificial neural net with human cerebral complexity ($O(10^{10})$ units (neurons) and $O(10^{14})$ synapses), not to mention our limited understanding of the wiring necessary to create the correlations, inhibition and binding which are central to the information field theory of consciousness presented here. Yet we may imagine that in time we shall have the technological capacity to create such machine. Will it be conscious? And how will we know? Currently we cannot even answer such questions about our colleagues.

The artificial neural network with its dynamic synapses need not be the only model of a plastic processing system (a machine) which can induce an information field which is dual to its primal processing capabilities. What of the digital computer itself on which runs a simulation of an appropriate plastic processor? Will it generate a dual information field and be conscious? Of course, we recognize this question as referring to an information field augmentation of the principle of strong AI.

Measurement and falsifiability

In principle the constituents of the theory presented here can be measured, so that the theory itself is falsifiable. Sufficient is a wiring diagram of the brain and the ability to probe simultaneously the activity of an enormous number of synapses. Naturally, this observation is subject to the taxonomic critique (see the Introduction) that we will thereby be measuring yet another neural correlate (albeit with its dual field construct appended) to consciousness and not consciousness itself. However, provided that we understand the brain as circuitry, this theory will predict what scene is being experienced by the possessor of a brain as a result of such measurements. The theory can thus be verified (falsified) by asking the possessor a simple question!

Appendix

Here we give details of the comparison of $\bar{I}$ with v^e in the cases $n = 1$ and 2.

(i) Consider first the scalar case, $n = 1$. Here $v^a = 0$ or 1.

If $v^a = 0$, we have

$$v^e = h(-\theta) = 0,$$

$$\bar{I} \equiv I = h(-\theta) \wedge 0 = 0.$$

If $v^a = 1$, we have

$$v^e = h(s-\theta) = \begin{cases} 1, & s-\theta > 0, \\ 0, & s-\theta < 0. \end{cases}$$

$$\bar{I} = h(s-\theta) \wedge 1 = v^e.$$

So in the case $n = 1$, $\bar{I} \equiv v^e$.

(ii) Consider the case $n = 2$, where there are four possibilities $v^a = \binom{0}{0}, \binom{0}{1}, \binom{1}{0}, \binom{1}{1}$.

If $v^a = \binom{0}{0}$,

$$v^e = h(-\theta) = 0,$$

$$I = h(-\theta) \wedge \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\bar{I} = 0.$$

If $v^a = \binom{0}{1}$,

$$v^e = h(s_2-\theta) = \begin{cases} 1, & s_2-\theta > 0, \\ 0, & s_2-\theta < 0. \end{cases}$$

$$I = h(s_2-\theta) \wedge \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{cases} \binom{0}{1}, & s_2-\theta > 0, \\ \binom{0}{0}, & s_2-\theta < 0. \end{cases}$$

$$\bar{I} = \begin{cases} \frac{1}{2}, & s_2-\theta > 0, \\ 0, & s_2-\theta < 0. \end{cases}$$

If $v^a = \binom{1}{0}$, we have the results of the case $v^a = \binom{0}{1}$, with s_1 replacing s_2.

If $v^a = \binom{1}{1}$,

$$v^e = h(s_1+s_2-\theta) = \begin{cases} 1, & s_1+s_2-\theta > 0, \\ 0, & s_1+s_2-\theta < 0. \end{cases}$$

$$I = h(s_1+s_2-\theta) \wedge \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{cases} \binom{1}{1}, & s_1+s_2-\theta > 0, \\ \binom{0}{0}, & s_1+s_2-\theta < 0. \end{cases}$$

$$\bar{I} = \begin{cases} 1, & s_1+s_2-\theta > 0, \\ 0, & s_1+s_2-\theta < 0. \end{cases}$$

(Compare the inequality conditions appearing several times in this appendix with the dendrital aborization hypothesis of Willner et al, 1995, the latter being critical for the self-organization (the emergence) of the locomotive oscillator in the mammalian spinal cord.)

References

Chalmers, D.J. (1995), 'Facing up to the problem of consciousness,' JCS, **3**, pp. 386–401.

Chalmers, D.J. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press).

Churchland, Patricia S. (1996), 'The hornswoggle problem,' JCS, **3**, pp. 402–8.

Corben, H.C. and Stehle, P. (1950), *Classical Mechanics*, (New York, John Wiley).

Crick, F. and Koch, C. (1995), 'Why neuroscience may be able to explain consciousness,' Scientific American **273**, pp. 66–77.

Hameroff, S. and Penrose, R. (1996), 'Conscious events as orchestrated time-space selections,' JCS, **3**, pp. 36–53.

Hardcastle, V.G. (1996), 'The why of consciousness: A non-issue for materialists,' JCS, **3**, pp. 7–13.

Haykin, S. (1994), *Neural Networks , a Comprehensive Foundation* (NY: Macmillan).

Hut, P. and Shepard, R. (1996), 'Turning the hard problem upside down and sideways,' JCS, **3**, pp. 313–29.

Kairiss, E. and Miranker, W.L. (1997), 'Cortical Memory Dynamics,' Bio. Cyb., in press.

Linsker, R. (1988), 'Self-organization in a perceptual network,' Computer **21**, pp. 105–117.

Miranker, W.L. (1997), 'Interference effects in computation,' SIAM Review, in press.

Penrose, R. (1989), *The Emperor's New Mind* (New York: Oxford University Press).

Penrose, R. (1994), *Shadows of the Mind* (New York: Oxford University Press).

Russel, B. (1927), *The Analysis of Matter* (London: Kegan Paul).

Stapp, H. (1996), 'The hard problem: a quantum approach,' **3**, pp. 194–210.

von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics* (Princeton: Princeton University Press).

Wigner, E. (1961), 'Remarks on the mind-body problem', in *The Scientist Speculates*, ed. I.J. Good (London: Heineman).

Willner, B., Miranker, W.L., Lu, C. (1995), 'Self-organization of the locomotive oscillator,' J. Math. Biol., vol. pp.