Consciousness is an Information State
Induced by Hebbian Dynamics

Willard L. Miranker

YALE/DCS/RR-1136
August 1997
(Revision)

Consciousness is an Information State [1]

Induced by Hebbian Dynamics

by

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

Abstract: We introduce an information state associated with the action potentials, the latter encoding conventional unconscious neural processing. We show that the state generates a dual representation of the neural processing which mirrors the information conveyed by the neural circuitry itself. The information state is shown to be an indicatrix of consciousness, a property which is verifiable by experiment. This leads to the claim that the information state, the indicatrix itself, is the conscious experience of neural processing. We start with the Hebbian synapse whose dynamics are interpreted as an atom of awareness, and which is quantified in terms of the time rate of change of synaptic strength. We show how the information state is built up out of such atomic constituents. The mathematical development shows that consciousness arises through a coupling of internal (the dual information state) 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. Our model offers an explanation of the alternating experiences of so-called illusions. An explanation for the fitness advantage of consciousness in evolution comes as a by-product of our information theoretic approach. The relevance of our model to the issues of nonhuman consciousness, both animal and machine, is described.

[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. The manuscript was revised during a stay at the Mathematics Department, Stanford University in June 1997. The author is grateful to J. Keller for helping to make this stay possible and also for helpful comments.

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

1 Introduction

We propose that *conscious experience* corresponds to an *information state* which accompanies neural processing. The state is associated with, but is different from, the action potentials in terms of which conventional neural processing is conducted. The information state varies in strength, depending upon details of the neural processing, and when it approaches its (normalized) maximum value, the information state is the conscious experience of the neural processing which it parallels. Thus our approach is a coupling of internal (the information state) and external (the neural processing) properties of matter (B. Russell, 1927). There is a contrasting 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. In the presentation here, the causal agent is a physical threshold effect, and it is *consciousness itself* which emerges.

We start by reviewing the standard Hebbian synapse, and we interpret its dynamics as a *quantum or atom of awareness.* This awareness is expressed as information which is measured in terms of $\dot{s}(\equiv \frac{ds}{dt})$, where s is the synaptic strength. The information state, to be denoted by $\mathbf{I}$, will be shown to be an indicatrix of consciousness, a claim that can be verified by measurement. (To this extent, at least, the theory presented here is falsifiable.) $\mathbf{I}$ is supported by a collection of neurons, and the value of $\mathbf{I}$ at any one of those neurons is a function of the information contained in that neuron's set of afferent synapses. We take the value of $\mathbf{I}$ at the neuron to be the average over this set. We show that this average (that is, the information state's value) is connected 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 state is an exact correspondent of the unconscious signals being conveyed and processed by a collection of neurons in the customary sense.

Since the information state 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 the information state introduced here, this indicatrix of consciousness, is consciousness itself.

Our model offers an explanation of illusions, alternating experiences corresponding to a fixed sensory input. This results from the introduction of a hysteresis effect into the model. Also, a by-product of our approach is an explanation of the fitness advantage of consciousness in evolution. This results from the use of an information theoretic notion applied to a collection of neurons, focusing on the latter as the generator of the information (state). Our model impacts the questions of nonhuman consciousness; animal and also machine.

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 momentum).

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.L. Miranker, 1997b 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. While the presentations made here seem to belong to the mysterian class of theories of consciousness, it is possible to argue that they belong as well to each of the other two categories: type A materialism and type B materialism.

In Section 2, we motivate our approach with a discussion of atomic awareness at a metamorphic level. Then we introduce the information state **I** 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 postulate that **I** is consciousness itself. This develops from (a) a primal/dual interpretation of neural processing/**I**, (b) the interpretation of **I** as an indicatrix of consciousness, and (c) an experimental method for demonstrating (b). In Section 4 we connect the model with several properties of consciousness: (i) contrasting the consciousness duality of Section 3 with a duality in quantum mechanics in which consciousness itself enters, (ii) degrees of consciousness, and (iii) binding. In Section 5, we apply the model to explain certain features of consciousness: (i) an explanation of the alternating experiences of so-called illusions and (ii) an information theoretic property of consciousness (of the indicatrix) with its Darwinian implication. Finally in Section 6, we indicate how our model impacts questions of nonhuman consciousness, first animal and then machine.

2 The Information State

For reasons of clarity, we shall deal with a simple model, the McCulloch–Pitts neuron with n input synapses. (In circuitry terminology, n is the fan-in, equivalently, the fan-out of the circuit elements: the neurons.) Then let v^e, a binary scalar, be the efferent neural activity of such a neuron. Let $v^a = (v_1^a, \ldots, v_n^a)^T$, a vector of binary scalars, be the afferent neural activity of that neuron, and let $s = (s_1, \ldots, s_n)^T$ be the real valued vector of corresponding synaptic strengths. The Hebbian synaptic dynamics

for each neuron are written as the following system of n differential equations,

$$\frac{ds}{dt} = H(v^e, v^a).$$

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

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

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

and $\theta > 0$ is a threshold.

From the differential equation, we see that s evolves in time. From the neural output equation, we see that the same is then true for v^e. The afferents v^a for one neuron are composed of the efferents of a collection of other neurons, so that the v^a are likewise time dependent.

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, so that the Hebbian dynamics are taken as a postulated and unreducible property of nature (in category, analogous to the law of gravity, say).[3] We say that the synapse 'experiences' v^a and v^e, and the nature of the experience is a tendency to change the value of s. We regard $\dot{s}$ as the potentiator of experience (in the synapse).

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 a tendency 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. We see that $\dot{r}$

[3]This characterization motivates a metaphysical view of the Hebbian dynamics. Should these dynamics become expressible in terms of underlying processes in the future, we should apply the words 'metaphysical' and 'unreducible' alternatively to those underlying processes as appropriate, leaving this presentation otherwise essentially unchanged.

potentiates the experience of the pair of particles. Viewed in this context our synaptic, information-state approach is a coupling of internal and external aspects of matter (B. Russell, 1927).

Specification of the information state

We seek to express consciousness in terms of a time varying normalized information state which we shall denote by **I** and which corresponds to a collection of neurons. **I** is taken to be a vector with a component corresponding to each neuron in the collection, each such component to take values in the interval $[0, 1]$. Motivated by the discussion in the preceding subsection ("The metaphor of experience"), we make the following postulate.

> *Each component $\dot{s}_j$ of every neuron's vector $\dot{s}$ is an indicator of an atomic awareness associated with the corresponding afferent synapse. (That is, $\dot{s}_j$ is the value of a consciousness indicatrix for the j-th afferent synapse.)*

Next let I denote the component of **I** associated with a particular neuron. Then we further postulate that

$$I \textit{ is a function of that neuron's vector } \dot{s}, \textit{ viz } I = I(\dot{s}).$$

In primitive living systems, awareness[4] is sometimes viewed as a simple attraction or repulsion[5] (compare also Keplerian motion). So for this reason, and for reasons of convenience in the presentation here as well, we shall specialize the function $I(\dot{s})$ to a function $I(sig\dot{s})$. The value of $I(sig\dot{s})$ is to be built up out of the components $sig\dot{s}_j$ of the vector $sig\dot{s}$. For this purpose, we introduce a vector σ with binary components σ_j as follows.

$$\sigma_j = \frac{1 + \text{sig}\dot{s}_j}{2},$$

where $sig0 \equiv -1$. σ_j takes the value $1/0$ corresponding to the postulated attraction/repulsion in the j-th afferent synapse,[6] that is, corresponding to the signature of $\dot{s}$.

Hebb's idea that the strength of a synapse, s_j, increases/decreases if the afferent and efferent voltages, v_j^a and v^e, agree/disagree may be embodied in the differential equation for the synaptic dynamics. This will be the case if the Hebb function $H(v^e, v^a)$ is chosen so that the triples $(v_j^a, v^e, \dot{s}_j)$ take values corresponding only to

[4] Awareness in primitive living systems is commonly viewed as unconscious.

[5] Compare the swimming motion of E. coli in the direction of increase of nutrient concentration (the attraction), or the jumping of the bacterium to a random location when an adequate increase of nutrient direction is not available (the repulsion).

[6] Returning briefly to the Keplerian metaphor, we could identify the separate attractive and repulsive sensations there with the two values $(1 + \text{sig}\ \dot{r})/2 = 1$ or 0. Indeed this suggests the terminology *sensation* for σ.

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

Table 1: excitatory case ($j \in \{+\}$)

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

Table 2: inhibitory case ($j \in \{-\}$)

entries of the first three columns of Tables 1 and 2. The first table corresponds to the case that the afferent synapse receives excitatory input and the second table to the corresponding inhibitory situation. We denote the collection of indices j in the case of excitatory/inhibitory afferents as $\{+\}/\{-\}$. That is, $j \in \{+\}/\{-\}$ if the j-th synapse is a receiver of excitatory/inhibitory input. Also let n_+/n_- denote the number of indices in $\{+\}/\{-\}$. That is, $n_{+/-}$ denotes the number of afferents which receive excitory/inhibitory input. (Recall that $n_+ + n_-$ may be as large as 10^5 in the human brain.)

Note that we may write σ as the Boolean 'and' function of v^a and v^e, where componentwise

$$\sigma_j = \begin{cases} v_j^a \wedge v^e, & \text{if } j \in \{+\}, \\ v_j^a \wedge \tilde{v}^e, & \text{if } j \in \{-\}. \end{cases}$$

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

Now let $\Sigma_{+/-}$ denote the sum $\Sigma_{(j\in\{+\})/(j\in\{-\})}$, and let I be defined as the following signed average of the σ_j.

$$2I = 1 + \frac{1}{n_+}\Sigma_+\sigma_j - \frac{1}{n_-}\Sigma_-\sigma_j.$$

A deeper understanding of these ideas might indicate a more complex, perhaps nonlinear function of the σ_j be taken to specify I. 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. Similarly let $\|v^a\|_{+/-}$ denote the number of excitatory/inhibitory afferents which are firing.

We take the value of the information associated with the neuron to be I. Since

$v^e = 1$ or 0, we see that

$$2I = \begin{cases} 1 + \frac{1}{n_+}\Sigma_+ v_j^a, & v^e = 1, \\ 1 - \frac{1}{n_-}\Sigma_- v_j^a, & v^e = 0, \end{cases}$$

$$= \begin{cases} 1 + \frac{1}{n_+}\|v^a\|_+, & v^e = 1, \\ 1 - \frac{1}{n_-}\|v^a\|_-, & v^e = 0, \end{cases}$$

with $I \rightarrow v^e$ (= 1 or 0), increasing or decreasing, as the case may be, as the correlation among the afferent synaptic activities of the neuron increases. In particular, the information I is zero if the neuron does not fire, and it is equal to the specific input intensity if the neuron does fire. (Note that this is a *threshold effect* for I. Thus the (normalized) information quantity I mirrors v^e. In the subsection on illusions in Section 5, we shall suggest a reason for introducing hysteresis into this threshold effect.)

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 normalized information values comprise a state **I** supported by such large collections. What we have derived is the state's value I at one position, at one neuron. Moreover, since I mirrors v^e, **I** mirrors the actual unconscious information encoded by the action potentials in the corresponding neural collection.

Gain and emergence

As noted, the threshold effect (i.e., gain) for the generation of v^e is carried over to I by this definition. In Section 3 we shall interpret the state **I** as consciousness, and so, we recognize this threshold behavior as the characteristic *emergence quality* of consciousness. The threshold value will be denoted by φ, which equals the value of $\|v^a\|_+/n_+ - \|v^a\|_-/n_-$ at which emergence occurs.

Field of information

In an earlier version of these ideas (W.L. Miranker, 1997a), **I** was called an information field because it is a quantity distributed over a spatial collection of neurons and because of its metaphoric analogy to gravity. Since fields in physics are usually functions of continuous variables, **I** has been given the more formally correct descriptor of state in this presentation.

3 Consciousness and Quale

Motivated by the development in Section 2, we shall introduce a hypothesis identifying **I** as consciousness (i.e., as conscious experience). The steps leading to this are:

a) the development of a primal/dual interpretation of (neural processing)/(the information in the state **I**).

b) the interpretation of **I** as an indicatrix of consciousness.

c) an experimental method for demonstrating (b).

Scenes, duality, and the indicatrix

We have seen that the value of 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 call *scene* the physical information (a color, a sound, an odor, a pain, ...) which is at any instant of time being processed by a collection of neurons. This scene is encoded, i.e., is represented by the neural activity (the action potentials) 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 state **I** is an alternate, a *dual representation* (a dual encoding) of the scene. The postulate made in Section 2, that **I** is an indicatrix of consciousness, can be verified by measurement, a task to which we now turn.

Measurement and falsifiability

In principle, the constituents of the theory presented here can be measured, so that the theory itself is falsifiable. That is, the postulate that **I** is an indicatrix of consciousness can be demonstrated by experiment. A wiring diagram of the brain and the ability to probe simultaneously the activity of an enormous number of synapses is sufficient. If we understand the brain as unconscious circuitry, this theory will predict what scene is being consciously 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!

The consciousness hypothesis and quale

Formally we may identify a property with its indicatrix. So in particular, we make the following hypothesis.

The information state **I** *is consciousness.*

We are all the more motivated to make this identification, because **I** is built up out of atomic awareness at the synaptic level. We expressed this awareness as experience in a metaphoric sense. Yet at a larger scale, we perceive a gap between **I** and conscious experience as we know it personally. The gap exists in perception only, because the explanation of consciousness offered here is not a reductive one. Thus the gap can be bridged by constructs of a metaphoric sort, perhaps an elaboration of our procedure at the synaptic level. (This could be similar to the filling of perceptual gaps in our understanding of other unreducible properties of nature, such as time, space or gravity.)

The consciousness hypothesis having been made, we may use the term qualia for the dual version of a scene.

4 Properties

We now give interpretive comments connecting the construct of the model with several properties of consciousness. These deal with (i) contrasting the consciousness duality developed here with a duality in quantum mechanics in which consciousness itself plays a role, (ii) degree of consciousness, and (iii) binding, the connecting of different kinds of sensory information into a single conscious experience.

Duality and quantum mechanics

Let us contrast the consciousness/unconsciousness duality (i.e., the 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 according to which these waves are propagated. 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 state **I**. For us the causal agent is a physical effect, and (as we have hypothesized) consciousness itself emerges as the information state.

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 our model. 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 among the afferents, the consciousness which is weakening disappears altogether. Compare this with the comment on gain near the conclusion of Section 2.

Binding

This weakening and 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) correlations 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 quality 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.

5 Applications

We apply the model to explain two functional features of consciousness: (i) the alternation of experiences characteristic of illusions, and (ii) the development of the fitness advantage of consciousness in evolution.

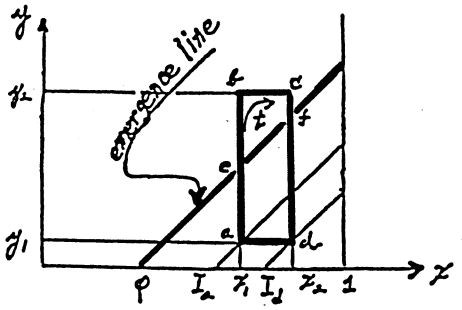

Figure 1: An illusion loop set over indicatrix level lines

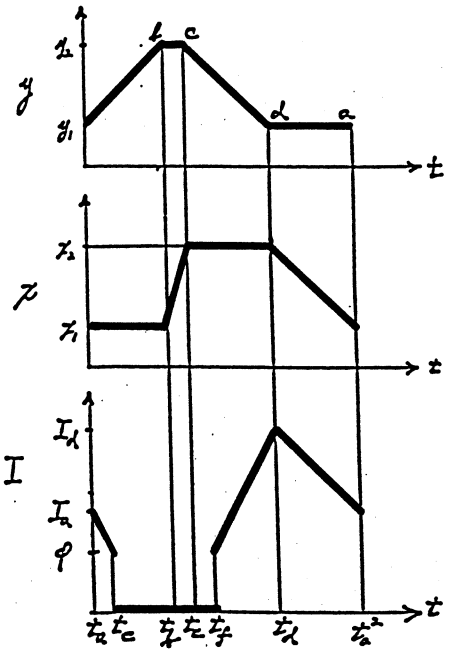

Figure 2: Cyclic excursions of x, y, and I around the illusion loop

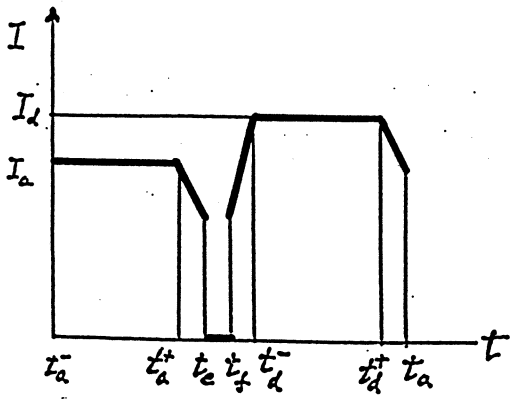

Figure 3: Dynamic excursion of the two illusion values I_a and I_d

Illusions and hysteresis

Conscious experience is subject to illusions. There are well-known examples where one specific input of an image, say, provides one conscious experience (interpretation) followed by a second. Sometimes these interpretations flip repetitiously. We shall show that our construct, the information state **I**, may be invoked to explain this phenomenon.

We consider a hysteresis effect in the response of I to the correlation/decorrelation of afferent activity in a neuron. The excursion around the hysteresis loop is induced by auxiliary inputs, alternately inhibitory and excitatory, to the neural processing of the scene in question. These inputs may arise from any of the many sources which bind to the processing of the current (illusory) scene(s) in question. The explanation, an alternation between two values of I, is illustrated in the figures.

In Figure 1, we plot contours of I versus $(x, y) \equiv (\|v^a\|_-/n_-, \|v^a\|_+/n_+)$. We also plot a putative dynamic excursion of (x, y), namely the loop (a, b, c, d), which is chosen to cross what we shall call the *emergence line*: $x - y = \varphi$. (φ is the threshold value for the emergence of I as described in Section 2, the subsection on "Gain and emergence.") Note that $I \equiv 0$ above this line, and it has its maximum value, unity, in the lower right hand corner of the figure. The points e and f are where the emergence line crosses the chosen loop. In Figure 1, the location of x_1, x_2, y_1, and y_2 (equivalently, the choice of the loop) are arbitrary except that they are chosen conveniently in order to illustrate our explanation of illusions. The indicatrix values I_a and I_d result from fixing the loop.

In Figure 2, we plot I, x, and y versus time. The values of time corresponding to passage through vertices of the loop are denoted $(t_a^k, t_b^k, t_c^k, t_d^k)$, where $k = 1, 2, \ldots$ indicates successive cycles around the loop. (For clarity we shall sometimes suppress the superscript k.)

In Figure 3, we plot a time scaled version of the I versus t part of Figure 2. In particular we have allowed this excursion around the loop to dwell at the point a/d over the time interval $[t_a^-, t_a^+]/[t_d^-, t_d^+]$. The corresponding values of I are I_a/I_d, and these values are the contributions of the neuron in question to the two different information state values of this illusion. The remaining time intervals of the excursion have been considerably reduced. As the loop is repeatedly traversed, the result is a repetitious and quick flipping between different experiences of the same scene, the duration of each experience and of each flip being arbitrary.

Infomax and evolution

The correlation which is central to the emergence of the state **I** (of consciousness) in our theory has an infomax interpretation which we shall describe, taking the case of no inhibitory connections, the latter for reasons of clarity. 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 $\mathcal{I}(v, x)$ 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 $\mathcal{I}(v,x)$ is associated with the information state **I**, 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, the components of which are different neuronal outputs. Here v_j^a is the output of a neuron j, $j = 1,\ldots,N$.

The average mutual information $\mathcal{I}(v_j^a,x)$ for each neuron in this collection is increased with redundancy in the (values of the) components of the vector v^a (under a set of conditions on the neuronal gain function, on the type and independence of the input noise, etc. See S. Haykin, 1994, pp. 455-8 for details. See R. Linsker, 1986 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 quality associated with increasing the mutual information in a network.

This suggests the fitness advantage of consciousness in evolution.

6 Nonhuman Consciousness

We comment on the connection of our model of consciousness, first to (nonhuman) animals and then to machines.

Animal consciousness

Let us denormalize the indicatrix **I** by deleting the factors $\frac{1}{n}$ in the equation defining I in Section 2. It is an assumption compatible with our approach here, that the consciousness of an animal will have a degree related to this size of the maximum I value. The maximum value of I is (proportional to) the fan-out.[8] Of course, nonlinear effects, especially in the form of threshold effects are possible and likely. Thus, our methods don't exclude consciousness from being qualitatively different from one species to the next.

Machine consciousness

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 that its processing elements have the capacity to induce or generate a dual system of information representation. This ability to generate a dual representation of information seems unlikely for the digital computer as it is

[8]Information concerning the taxonomic variation of fan-out is not available. The fan-out varies considerably among the types of neurons in the human brain, and this should have bearing on human consciousness and its 'location'. For example, the fan-out approaches unity in the midget bipolar system and in the climbing fiber system in the cerebellum. The granule cell-parallel fiber system has a fan-out of the order 10^2. (These observations were pointed out to me by G. Shepherd.)

currently constituted. (Recall that the basis for this capacity in the brain is $\dot{s}$, the Hebbian dynamics, and the postulated feature of atomic awareness in the synapses.) Artificial neural nets have plastic synapses (the latter were identified by us as the potentiators of awareness in the brain). So it may be possible to build such artificial nets 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, binding and hysteresis which are central to the information state 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?

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 state which is dual to its primal processing capabilities. What of the digital computer itself which runs a simulation of an appropriate plastic processor?[9] Will it generate a dual information state and be conscious? Of course, we recognize this question as referring to an information state augmentation of the principle of strong AI.

[9]For that matter, what of a digital computer which runs a simulation of the brain? Some mysterians, e.g. R. Penrose, say that such a state of affairs is a contradiction of terms; that a digital computer is at most capable of (Turing) computable functionality, whereas the mind (consciousness) is noncomputable. See W. Miranker, 1997b also.

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).

Chalmers, D.J. (1997), 'Moving forward on the problem of consciousness,' JCS, **4**, pp. 3–48.

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. (1986), 'From basic network principles to neural architecture,' PNAS (USA), **83**, pp. 7508–7512.

Miranker, W.L. (1997a), 'Consciousness is an Information Field Induced by Hebbian Dynamics', TICAM Report 97-11, June 1997, The University of Texas at Austin.

Miranker, W.L. (1997b), '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).

Russell, 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).