THE FRAGMENTATION OF THE UNIVERSE AND THE DEVOLUTION OF CONSCIOUSNESS

Stephen L. Thaler, Imagination Engines, Inc.
St. Louis, MO 63146-4331, USA. Email; sthaler@ix.netcom.com

From the U.S. Library of Congress

Registration No. TXU00775586, 1997

Abstract: Contrary to the popular notion that consciousness is the result of a noble evolutionary process, I speculate that this rather ill-defined concept and phenomenon may be the result of the fragmentation of an otherwise completely connected and totally 'feeling' universe. As various regions of this universe topologically pinch-off from the whole, connection-sparse boundaries form over which sporadic and impoverished information exchange takes place. Supplied with only scanty clues about the state of the external world, abundant internal chaos drives these small parallel processing islands into multiple 'interpretations' of the environment in a process we identify with perception. With further division of these regions by insulating partitions, the resulting subregions activate to lend multiple interpretation to the random activations of others in a manner reminiscent of internal imagery. The spontaneous invention of significance by this weakly coupled assembly of simple computational units to its own overall collective behavior is what we have grown to recognize as biological consciousness. We thereby come to view human cortical activity as a highly degraded approximation to the original and prototypical cosmic connectivity.

I: Introduction

At the heart of the debate over the nature of consciousness, is the inherent limitation of human model formation. Whether engaged in scientific, philosophical, or even religious theorizing, the conceptual primitives forming the basis of the most rigorous of models have very little absolute validity. They are nothing more than analogies whose appeal critically depends on their neurological habituation. With sufficient scrutiny, we may always reveal some level at which the model's chain of associations begins and the primitives involved lack foundation. For the much celebrated reductionist physics, for example, fundamental notions of mass, charge, space, and time can only be defined by analogy with everyday notions at the human level such as globs of matter (particles) and water disturbances (waves). From a theological perspective, all reality stems from a god figure modeled from our everyday experience of a human patriarch building his house and tending his children. Note that in both extremes, there is no success in absolute definition...only the creation of analogies that ring with familiarity. We squint at the reality, caught in an endless pursuit of ever more fundamental particles and surmising the god behind the god. Addressing these issues with theories that are circular, we inevitably refer to common concepts in our everyday world. Ultimately, there is no absolute explanation, only description based upon culturally embraced paradigms.

It is for this fundamental reason that theories about anything are born, promulgated, and die, independently of their veracity. With myriad analogies to choose from, views come and go, as the most sophisticated of rationales mutates with consensus. Such is definitely the case with theories of consciousness, where each of the scientific and philosophical factions involved possess entirely different experience and analogy bases to draw upon. The result is a veritable Tower of Babel, and a proliferation of diverse notions that render intercommunication among factions difficult, if not impossible.

This obvious fiasco, which at first seems daunting and insurmountable, may have a profound bearing upon our ultimate model of consciousness, clearly defining the underlying problem: **All we can expect from an analogy machine (i.e., the brain) is nothing more than an analogy about itself.** We should never raise

our expectations to the point of anticipating some all-encompassing truth about the mind or of consciousness. All that we may hope for is the evolution of pragmatic theories that provide us with varying degrees of (1) predictive power, (2) technological application, and perhaps (3) psychological comfort.

In this paper, we propose a new and useful analogy that might at first be called connectionist, but in retrospect may be deemed more of a "disconnectionist" picture of consciousness. In this portrayal, these isolated regions (to which I attribute consciousness) operate not by some glorious destiny or "New Age" quantum mechanical effects, but by a wide range of randomizing factors so prevalent and cheap within the cosmos. Although such a model may at first seem cold and bleak, it does achieve a good measure of all of the above-named advantages, including psychological haven. Furthermore, and most importantly, it provides for a quantitative reductionist model of consciousness yielding immediate technological pay off.

II: Fragmentation of the Universe

Consider a fully connected universe, U existing at some arbitrary time t_0. Without developing analogy-based theories about the nature of the fundamental interactions involved, we will simply assume that there exists an immense number of forces (i.e., connections) among all of its contained entities (i.e., nodes). Perturbation of any connection or node will result in the evolution of U into a series of progressive and distinguishable states, just one of which I show in Figure 1. Therefore, the slightest 'nudge' to any of its myriad masses or charges will result in the celebrated "butterfly effect" (J. Gleick, 1987) whereby this small perturbation propagates throughout the system, affecting the remotest of particles. Here, the rather simplified symbolism of black and gray nodes conveys the wide range of accessible position and momentum states available to each entity within this collective structure. Connections consist of physical interaction, encompassing the whole gamut of physical forces, ranging from gravitational to electrostatic to weak nuclear interactions.

Any portion of the fully connected U will respond to changing activity in any of its other parts. To anthropomorphize, any piece of this universe can 'feel' the 'experience' of any other piece of itself. Therefore, should a perturbation arise in some distant region of U, other segments will be aware, owing to the multitudinous real-time couplings between highly removed regions. Flooding this relatively static system with various forms of readily available noise and fluctuations, all regions of the system will evolve through their allowed states in response to perturbations within all other regions. In a sense, the system will visit a long series of impressions about itself. Identifying U with some prototypical physical universe, we form the mental construct of a universe that is attentive to itself to within the time delays imposed by special relativity.

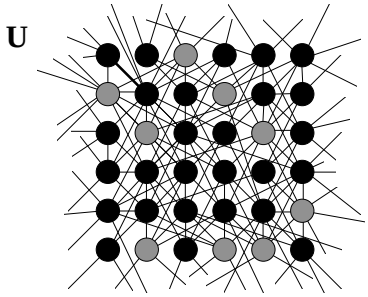


Figure 1. A connectionist universe, U. The myriad connection weights represent all physical forces present.

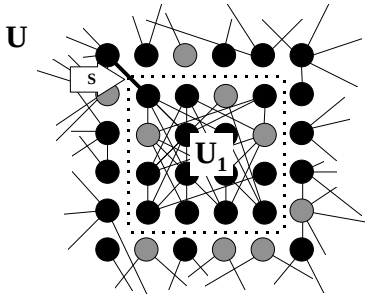


Figure 2. Dissolution of connections creates sub-universe, U_1 with a single surviving link, S.

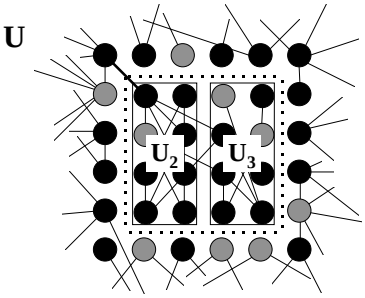


Figure 3. Further dissolution of connections creates further fragmentation into U_2 and U_3.

Since connections not only form within U, but also dissolve, it is quite possible that regions of U may become isolated over time. Only a few token connections may remain such as the single surviving link 'S' shown in Figure 2. Effectively, a separate universe, U_1 forms, rendered insensitive to the majority of activity in its outer world. Sporadic information may occasionally transmit across the gulf by remnant

channels such as S. U_1 must then evolve into some state that is consistent with communication across this single, intact connecting channel. However, because the majority of links, and hence constraint relations, are now absent, U_1 may now visit a whole manifold of states subject to the single bridging condition at S. Mediated by entropic factors similarly trapped in U_1 by the topological pinch off from its parent world, internal chaos will push U_1 through a series of potential states until it arrives at a particularly stable condition (i.e., it has located an *attractor basin*).

We note than within the prototypical universe, the collection of entities destined to become the cluster U_1, would confine themselves to a set of states entirely consistent with conditions beyond their boundary. Therefore, for every event in the external universe, this region would reside within some precise and narrow band of states that could serve as a label or classification for each of the external conditions. With imperfect connectivity, as is now the case, U_1 will visit a much broader range of inner states for any given outer condition. Because the band of states available to U_1 now takes on a many-to-one relationship with the external conditions, ambiguity arises. That is to say, an outside observer looking internally into U_1, would find it difficult to exactly surmise external activity. We therefore say that this connectionist cluster has found an 'interpretation' of its environment, in contrast to precisely defining its surroundings. This semi-isolated connectionist cluster thereby achieves a kind of 'perception' by means of fragmentary impressions about the external universe. Slow to gather in clues about external happenings, outer developments come as 'shocks' or 'surprises' to it.

Over time, more connection barriers may spontaneously form, leading to the partitioning of U_1 into the distinct regions U_2 and U_3, loosely coupled by the two surviving connection traces portrayed in Figure 3. Inspection of that figure reveals that U_2 is still in contact with the external universe by a single channel, while U_3 can only see activity within U_2. Therefore, while internal noise may drive U_2 through many interpretations of the external universe's states, U_3 can only activate to interpret the internal states of U_2. Again likening the situation to the human condition, we say that internal imagery (i.e., bad guesses at the state of the world) occur in U_2 and that U_3 may now interpret the content as well as the raw progression of that imagery.

The U_2-U_3 cluster now represents a totally passive, yet 'cognitive' entity that may imagine various alternative scenarios within its sensed neighborhood. U_3 will register any number of activation patterns in response to each of U_2's many excitation states. The U_2-U_3 conglomerate possesses a stream of 'thoughts' (i.e., its activations) about its environment and in turn, has thoughts about these thoughts in a crude form of metacognition. Furthermore, U_2-U_3 possesses a primitive sense of self-identity because of its physical detachment from all else. It still is impotent to affect its environment. It can only contemplate.

III: The Inevitable Survival and Proliferation of Fragments

If at any given time in the evolution of the universe, there exists a degree of connection homogeneity, C_0, then at some later time t, this homogeneity may decrease to a level $C_t<C_0$. That is, certain regions of U will become connection deficient, leading to the boundaries discussed above, while other regions will become more dense in interconnections. Such a trend is not the result of some kind of cosmic will. It is simply the statistical fact that fluctuations are inevitable.

Therefore, we should see the cumulative growth of isolated connectionist clusters such as the U_2-U_3 conglomerate, only at much finer degrees of subdivision, containing now N subclusters, U_1, U_2, ..., U_N. If the overall boundary between the external universe, U and the conglomerate U_i (i=1,..., N) is more connection-deficient than the wall between any of the U_i, then the composite structure will be more in communication with itself than with the external universe. In other words, it is numb to its environment and sensitive primarily to itself. Furthermore, since the conglomerate entity now contains N multiple agents, all loosely coupled, there is more chance for 'surprise' of the assembly, by the unexpected excitation of any of the U_i into some novel activation pattern (i.e., a rare distribution of position and momentum states).

If the inevitable now happens and chance mishaps occur within such assemblies, leading to the increased survival value of the connection sparse barrier surrounding these clusters, then we will see a proliferation of such like insular regions. Among such adaptation features will be the recruitment of subclusters into organs for interaction and manipulation of its immediate environment, limbs for locomotion, and internal organs for its internal management and reproduction. That is, sparing the usual Darwinistic discourse, the inorganic clusters have evolved into the more complex systems that our parents told us were 'alive' and 'organic.'

In surviving the encroachment of the environment, those biological organisms that can enrich their internal connections will endow themselves with the ability to anticipate external activity and hence threats to their existence (i.e., preservation of their boundaries). To achieve this end, the raw connectionistic forces present within the primordial universe may undergo a reshuffling from the various forces at a distance into an electrochemical system. One result we know as the primate species homo sapiens, within which roughly 100 billion cortical neurons and their approximately 100 trillion interconnections form a model of the external universe. The state of this neural assembly is free to independently evolve, providing various survival notions to its host. Just as in the case of the external world, evolution internal to this connectionist system becomes enabled through various random factors. In Figure 4-5, for instance, we see events unfolding in a highly idealized physical universe as chaotic forces eject a single coulombic charge. Subsequently this simple physical system reequilibrates into a new stable state representing a local minimum in potential energy or a *potential well*.

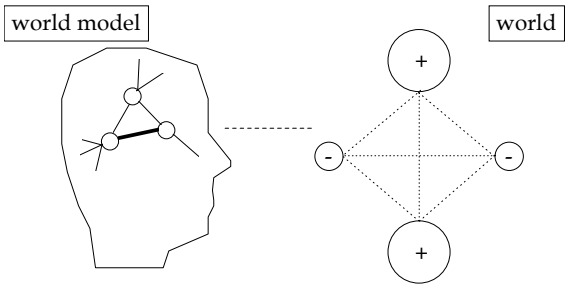


Figure 4. Electrochemical connections within cortex form a model of connection strengths within the external universe.

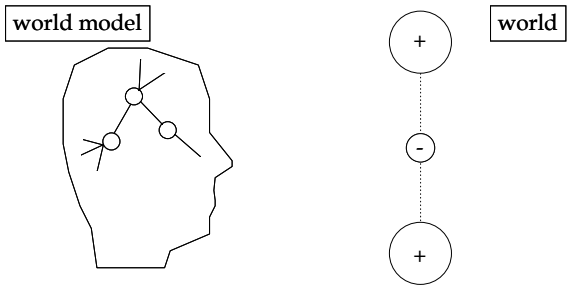


Figure 5. Internal electrochemical perturbations simulate perturbations in the external universe.

In similar manner, a particular pattern of neuronal activation within cortex may correspond to the internal image of the 'real-world' charge system of Figure 4. Neuronal chaos may transiently reduce or nullify a synaptic connection producing the internal image of the final state charge pattern in Figure 5. In complete analogy to the charge system, the cortex transitions between two network attractor basins that correspond identically with the two potential energy wells of the external coulombic system. Within the transition between these two basins, an inner dynamical system simulation evolves. Thus, a physcist may dream of this evolving charge system as the connections between neurons that store memories relax. In like fashion, the cortex is able to rehearse more complex scenarios within its environment, whether it be evaluating various danger scenarios, planning movement, or inventing various manipulations of the environment. In the same way, waking conscious imaginings of potential real-world entities and phenomena activate purely through internal noise factors.

IV: The Virtual Input Effect and Creativity Machine Paradigm

Inspired artificial intelligence

Therefore, we may view the cortex as a miniature approximation to the universe, evolving as a result of inner tumult to preview various scenarios and useful possibilities. Driven by this perspective, I have constructed relatively simple artificial neural network (ANN) models mimicking this process and yielding useful technological and esthetic results (Thaler, 1996a). Repeated implementation of this model to

numerous typically intractable problems has led to a number of astounding observations that support a chaos-driven model of consciousness.

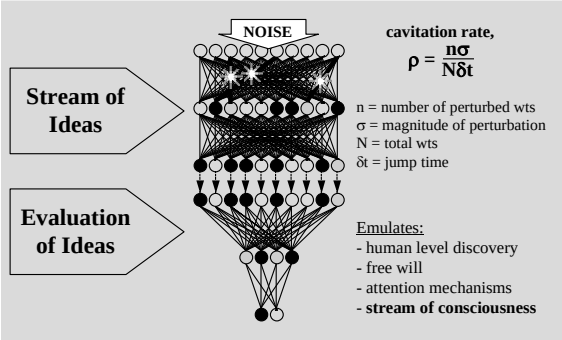


Figure 6. The Creativity Machine Paradigm

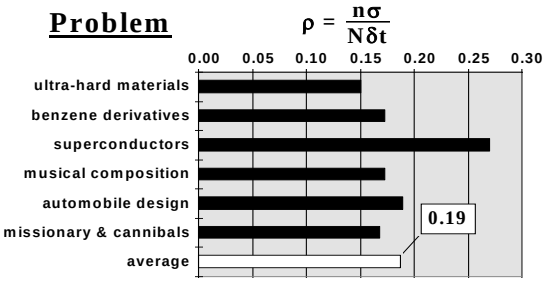


Figure 7. Optimal Cavitation Rate

In Figure 6, we see the essential components of the ANN model, two simple feedforward networks, trained by standard backpropagation technique paradigm (Rumelhart, Hinton & Williams, 1986). We imagine that the first net trains in some specific conceptual space, to relate any given pair of input and output vectors in that space. Likening these inputs to sensorium, this network is analogous to the subcluster U_2 depicted in Figure 3. The second net trains to associate any given output from the first to some other vectorialized characteristic or figure of merit. This second net generally corresponds to the network U_3, providing an interpretation to events witnessed in the former network. Purposely introducing internal noise or chaos into this first net will cause a series of network activations reminiscent of its training exemplars. It does so due to a process I call *internal vector completion*, whereby any pattern of internal disruption within a given network layer propagates to subsequent network layers. These downstream layers then interpret this spurious signal as some activation pattern the network has already seen in training. Viewed from the dynamical system point of view, the network passes through a series of stable states or attractor basins, each distinguished by some distinct pattern of neuronal activation. The internal noise or random events introduced simply act to 'kick' the net into any one of these many stable states. In the midst of this stochastic interrogation, the network activates as though it is sensing a stream of training input examples, when in fact, we may have arbitrarily clamped inputs at constant values (i.e., It is fixated upon unchanging sensory signals propagating along any number of sparse sensory linkages as S in Figure 2.). I have coined the phrase "virtual input effect" to convey the apparent stream of events occurring at the net's inputs where, in fact, there are none (Thaler, 1995).

Until now, we have been discussing mild internal perturbations to network architecture that largely preserve the network connection weights. That is, the random disruptions to the net's connection weights (or processing units) have been small enough to conserve the attractor basin 'landscape,' with each attractor representing an intact memory stored in the network. If we gradually increase the magnitude of internal perturbation, we begin to mutate this landscape via the partial destruction of connection weights. In the process we create whole new basins, corresponding to false memories or confabulations.

Of course, one might at first question the usefulness of such degraded memories. The pragmatic reality is that many of these confabulations represent novel concepts that the second, policing network may assess as useful. This, I speculate, is how totally novel thoughts arise as if from out of the blue, amounting to no more than corrupted memories. Because the large majority of weights involved in the mapping are intact, the emerging vectorialized thoughts are reminiscent of the concepts the net has 'seen' in training. These concepts therefore serve as extrapolations from the known body of knowledge. A stream of new and plausible concepts then emerges. It is now up to the second network in this "Creativity Machine" architecture to filter and segregate the most valuable of these. Viewed from the perspective of human semantics, the rules governing the conceptual space, and implicitly contained within the trained connection weights, soften to allow for new possibilities. The opportunistic network then seizes any deviant scenarios offering advantage.

I refer to this technique as the "Creativity Machine Paradigm" and quickly point out the dramatic advantages it offers to the world of artificial intelligence (AI), generating solutions to problems previously thought intractable. I have applied it to a wide spectrum of conceptual spaces, as represented by the small sampling of topics shown in Figure 7. Because each of the required component networks requires only examples rather than hard-won rules for training, we may construct technologically useful machines within hours.

In the case of very objective problems, as in the search for new ultrahard materials (Thaler, 1996a), we train the first net on "correct chemistry," exposing it to examples of known chemical compounds and phases. As a computer code introduces chaos to the connection weights of this net, chemical formulas emerge. Some of these compounds are rehashes of training exemplars, while other are totally novel from the network's perspective. The second net, trained to map these formulae to hardness, may now relay only the hardest materials to an archival file for later perusal.

For more subjective problems, as in musical composition (Thaler, 1994), we may expose the first net or "imagination engine" to sundry examples of accepted melodies. As the random perturbations infiltrate the net, new candidate melodies emerge, generally obeying the constraints dictating what statistically constitutes culturally accepted music. The second patrolling network or "alert associative center," trained by exposure to the likes and dislikes of a panel of musical aficionados, may now filter out only the most appealing songs from the conceptual stream, either playing them in real time, or storing them for later retrieval.

Regardless of the problem domain faced by the Creativity Machine, there is a preferred operating regime for the internal perturbation parameter I have come to call the *cavitation rate*, ρ (likening the sporadic increases and decreases in the discrete field of connection weights to the rarefactions and compressions in a continuous field of a boiling liquid). In Figure 7, we define this cavitation rate as

$$\rho = \frac{n\sigma}{N\delta t} , \qquad (1)$$

where n represents the total number of synapses affected in any given instant, σ the average magnitude of synaptic disturbances, N the total number of connection weights in the system, and δt, the time between shifts in this microscopic pattern of perturbation (after which the n affected connection weights are cyclically returned to their original trained values and the governing algorithm chooses new connection weights for perturbation). Referring to Figure 7, we see the optimized operating regime for a number of different Creativity Machine tasks (ranging from such standard problems in AI such as the 'Cannibal and Missionaries' dilemma, to the less conventional assignment of novel materials discovery) and note that there exists a relatively narrow band of cavitation rate, near a value of $\rho = 0.19$, suitable for generating an optimal number of useful concepts per unit time. I believe that within the construction of creative machines, as well as within neurobiology, this value is a useful and significant numerical constant. As we are about to see, this constant strongly suggests the equivalence between human cognition and the Creativity Machine Paradigm.

That there should be a preferred cavitation regime should come as no great surprise, when we begin to consider concept generation at the extremes surrounding this region. If for instance, we reduce the cavitation rate to nearly zero, there will be little if any turnover of network activation, and hence no ideas for the policing net to choose from. At the other extreme, a high cavitation rate would destroy the trained weight values contributing to the overall network mapping, and the policing net must then sift through myriad nonsense ideas in search of useful notions. The resulting trade-off regime, centered at $\rho = 0.19$, represents a gentle perturbation to a parallel processing system, allowing interrogation of any contained conceptual space. Mild confabulations of stored memories emerge, generating a family of slightly mutated yet advantageous ideas.

The emerging features of consciousness

I have described a strictly connectionist approach to generating useful concepts across a wide variety of problem domains. Peripheral to these pragmatic issues, and more pertinent to the central discussion, I believe that such connectionist systems exemplify some of the more salient features we tend to associate with the term "consciousness." For instance, the Creativity Machine has led to human-level discovery and synthesis, as in the recommendation of new ultrahard materials and in the generation of multitudes of original musical melodies. Therefore (much to the protests of speciesist within each of these disciplines) we regard this paradigm as an *intelligent agent*, producing human-level discoveries. Since these virtual machines have required only 'noise' as input, arriving at their own independent decisions and courses of action, such intelligent agents have demonstrated, in essence, 'free will.' Furthermore, such systems possess attentional mechanisms, whereby the policing net activates at the appearance of an acceptable output from its cavitating partner, in a manner reminiscent of the "ah-ah! I've got it." catharsis in human level discovery. (Quite typically, we equip such machines with supplementary focusing mechanisms, whereby the second net directs the neutralization of those processing units in the cavitating net generally not carrying their weight in the conceptual search process. Alternately, such focusing feedback may consist of the gradual nullification of internal noise when the imagining net appears to be on the 'right track.')

Most importantly, and central to this discussion, the Creativity Machine provides a working model of what we commonly regard as 'stream of consciousness.' That is, just as within our own inexorable turnover of images and thoughts, internal chaos relentlessly drives an artificial neural network through a succession of memories about the compact universe it has learned about in training (i.e., its training set). As the amplitude of noise in the system increases, the cyclically damaged ANN fails in reproducing intact memories and, in turn, activates into confabulation states - 'could-be' possibilities generally conforming to the general features of information in its training set. This succession of memories and of ideas, respectively, persists as long as there is chaos within the system.

Generalizing and enlarging this paradigm to neurobiological dimensions and complexities, it is now quite plausible that ever present noise within neurobiology (i.e., cross-talk between neural networks, fluctuations in cell membrane potentials, quantum mechanical tunneling of neurotransmitters across synapses, diffusing neuromodulators and hormones, etc.) may similarly 'kick' cortical networks into a sequence of states representing stored memories and, if the chaos is sufficiently intense, into ideas having various degrees of personal or sociological novelty. Such memories and ideas now span a much greater range of sensory modalities than in the case of the simple ANN, now bracketing a wide variety of virtual impressions including visual, auditory, proprioceptive, and abstract imagery. That there appears to be more semblance of logic, rather than one helter-skelter thought after another, is that once a spontaneous idea nucleates (i.e., the notion of some tempting food), various associative neural network chains causally activate in response (i.e., a favorable recollection of that food's taste, and requisite actions to procure that food). At any moment, this dominant activation pattern may fade, owing to the *refractory* nature of the neural apparatus, and be supplanted either by another association or some completely irrelevant notion emerging from the synaptic chaos.

For those subscribing to notions of explainability in nature, this concept of an inexorable noise-activated parade of internal imagery, or stream of consciousness if you will, is a most parsimonious theory. After all, it requires only omnipresent noise, combined with plentiful forms of static neurobiology. It describes how ideas and feelings originate, ostensibly from out of nowhere, without the necessity of sensory inputs. Furthermore, it mitigates the role of perception in defining consciousness, since sensoria are extraneous, aside from learning and manipulation, to the stream and its origins (i.e., we are still conscious within a sensory deprivation tank).

The equation governing stream of consciousness

Apart from the above 'Occamy,' there is a much more compelling correspondence with cognitive observation: The Creativity Machine Paradigm provides a quantitative model of stream of consciousness. It does so on two counts (1) in specifying how new thoughts arise by degradation of vectorialized brain

states, as previously described, and (2) by specifying the exact temporal pattern of consciousness. In the latter vein, recent investigations have shown that totally independent of the exact ANN architectural and functional details, Creativity Machines generate memories and ideas at rhythms identical to those of human test subjects (Thaler, 1996b). Representing both artificial and human consciousness streams by two distinct factors, D_0, a clustering parameter (i.e., the fractal dimension of the cognitive stream, Voss, 1988), and the total time Δt required to originate a predetermined number of thoughts, N, a simple chaotic, neurodynamic calculation yields the relation,

$$\rho = \Delta t^{-D0}, \tag{2}$$

where ρ is the cavitation rate defined above. Setting ρ equal to the value 0f 0.19, representing the optimal regime of Creativity Machine operation, results of the ANN simulations and human cognitive experiments fall into direct alignment with one another (Thaler, 1996b). Furthermore we see a trade off which rings with intuitive familiarity. That is, for human cognitive and Creativity Machine tasks alike, the left side of Equation 2 is a constant requiring that shorter task durations possess higher fractal dimensions (lower clustering) while more time consuming assignments occur at lowered fractal dimension (increased clustering). Translating into more instinctive terms, straightforward tasks such as naming 20 numbers as quickly as possible (i.e., counting) occur nearly linearly. More demanding tasks, requiring more creativity occur in a more huddled pattern. As subjective observers, we typically interpret such results as the varying degrees of confidence and topic familiarity of a speaker.

Within more detailed ANN studies we establish the source of the dichotomy that leads to either of these extremes in cognitive event clustering. Generally, a multitude of combinations of n and σ may result in the constant value of ρ in Equation 1. We find that in the case of large n and small σ (many small perturbations randomly spread among the connection weights with a resulting high fractal dimension, D_0) we see the evolution of straightforward, intact memories. In contrast, for small n and large σ (large synaptic disturbances inflicted upon just a few connection weights with a resulting small fractal dimension, D_0) the network as a whole has difficulty in classifying the internal perturbation as some known environmental feature and, as a result, activates into corrupted memories or confabulations. Some of these notions may qualify as not only novel, but possibly useful. Identically, we see the same dichotomy in the human cognitive results, with non-inventive memory search possessing high fractal dimension approaching 1, and the more challenging inventive tasks occurring at much lower fractal dimensions near 0.2.

From a physicist's point of view, the quantitative theory outlined above would be a sufficient model of the conscious stream, invoking the equivalent of a energy-time representation for this dynamical system. The energetics would correspond to the stability of various states, or memories, as measured by the widths and the depths of their attractor basins. (The emergence of memories would then correspond to the analogy of a marble flitting from pocket to pocket within an agitated muffin pan, as depicted in the side bar. Novel idea formation would then imitate the melting and deformation of the pan to create new pockets.) The time evolution of thoughts, irrespective of the detailed nature of these ideas, would be statistically specified by Equation 2.

From the standpoint of consciousness studies, perhaps the most important feature manifested by this system is the continuous chain of vectorialized concepts nucleated by the presence of chaos within a cavitating network. If we generalize this parade of thoughts to a multitude of sensory modalities, we gain a plausible mechanism for the simulation of the human stream of consciousness. I believe that it is this inexorable progression of internal impressions that is central to the matter, in turn driving all the other metacognitive feelings we normally associate with both cognition and consciousness. This wellspring of notions appears driven purely by internal noise. Therefore, to achieve a complete model of consciousness, we need to supply some neurodynamic mechanism for the production of subjective feelings (i.e., the 'hard problem') about this spontaneous stream. When we do so, we supply a model for the so-called 'buzz' (D. Chalmers, 1996) that distinguishes human from 'zombie.'

V: Subjective Interpretation of the Stream

Content nucleated associative loops

I have claimed that at the core of consciousness is an inexorable chain of internal imagery driven by the random forces present within neurobiology. Now I propose that all other features of consciousness emerge as surrounding neural networks (such as the semi-isolated connectionist cluster, U_3) interpret this stream of consciousness, likewise assisted by abundant neurobiological chaos.

Edelman (1989) has postulated that such neural turmoil is the primary driver of human perception, with Freeman (1990) subsequently corroborating this hypothesis by rigorous experimentation. I now exemplify and extend their thinking in Figure 8, where we see a very simplified human cortex sensing environmental features within a greatly oversimplified universe (i.e., consisting of the star, arrow, and box objects). The visual sensory organ is reminiscent of the single vestigial connection channel S in Figure 2, transmitting a relatively straightforward retinotopic image to a simple visual cortex. This raw image is in turn parceled out to the myriad, communicating neural network groupings, each of these distributed cascades of neurons, representing some concept or feeling.

Of course none of these subsequent associative neural cascades is quiescent. Background noise bathes them, thereby modulating their connection strengths with the raw image in visual cortex. In the language of *dynamical systems*, the various attractor basins of these neural groupings are expanding and contracting, pigeon-holing the image and thereby attaching different interpretations to the impressed image. It is as though a competition has emerged among all possible associated concepts and feelings, with elements of chance now introduced. An allegorical roulette wheel spins to determine which feelings and associations will activate. However, in this 'fixed' casino of neurobiology we find that the more strongly habituated concepts are more often the winners, activating to attach various kinds and degrees of significance to the raw perception.

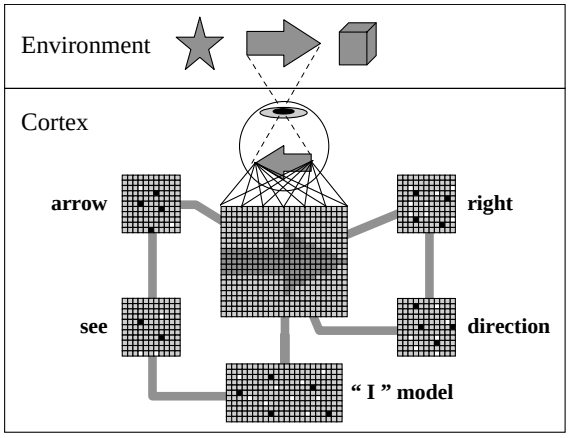


Figure 8. Neural Darwinism drives the interpretation of sensation (perception).

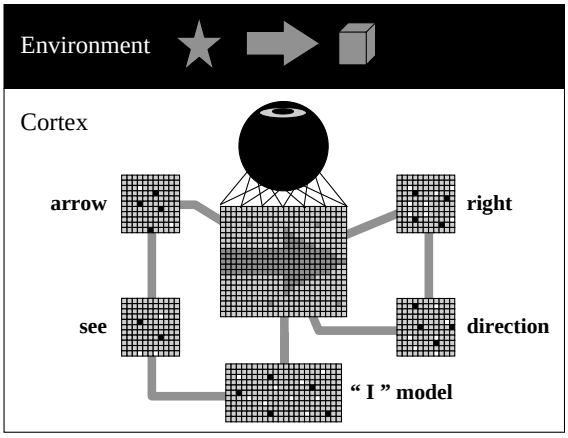


Figure 9. Likewise, Neural Darwinism drives the interpretation of noise-activated internal imagery.

In Figure 8, for instance, we imagine a situation in which the sensory organ is foveating upon the arrow feature within its environment. Presupposing that the concept of "right" is a strongly habituated concept, it now activates, in turn enlisting the similarly burnt in concept of "direction." Sooner or later some subsequent affiliation will link to a concept that refers back to the original seed of the chain, namely the retinotopic map of the arrow. An associative loop is thereby born.[1] In like manner, more introspective

[1] Note that the important issue of *binding* may be readily described by such looping associations. That is, if oscillatory elements are embedded in such a neural circuit, only those Fourier components or harmonics

loops may nucleate as when the "I" model (a neural grouping containing all self-perceptions), the concepts of "seeing" and "arrow." The resulting metacognitive thought is one of ourselves looking at an arrow. These and myriad other associative chains may nucleate, fade, and be supplanted by others over a period of seconds, in a competitive process coined "Neural Darwinism" by Edeleman (see, for instance, Franklin, 1995)

Remembering that the very simple visual cortex I have described is a neural network saturated in noise, any one of the previously observed environmental images (star, arrow, and box) should emerge through the *virtual input effect* previously described. Furthermore, because all the downstream associative networks have no mechanism for distinguishing between a real sensory input (i.e., perception) and noise-activated internal imagery within this visual cortex, nearly identical associative loops may activate (Figure 9). Therefore, in a manner similar to perception, surrounding neural groupings may join the Darwinian struggle to attach significance to internal images nucleating from noise. This activation process has resulted from the content of the visual cortex at a given instant in time. Consider now what happens as the visual cortex imagines a succession of all possible environmental features.

Associations driven by the raw succession of internal imagery

Commonly accepted among those who train and construct artificial neural network cascades is the accepted principle that anything may be mapped to anything else, given sufficient processing units and synaptic interconnections. For instance, we may train a musical ANN to 'listen' to the "Star-Spangled Banner" and instantaneously transform it to the Police tune, "Every Breath You Take." In the language of ANN theory, such a net has "memorized" this mapping. In like manner, we may apply the same connectionist model to neurobiology where many such mappings, between unlikely concepts, has arisen from eons of evolutionary demands. For instance, most mammals have a hard-wired neural map which relates the inherently unstimulating notion of 'concave meat slamming against convex meat' to the epitome of pleasure. The result is the sexual illusion that leads to the numerical robustness of the species. My strong suspicion is that this neural network illusion is just one of many more in nature's repertoire, including the major deceptions of consciousness that spawn feelings of distinctness, self-worth and self-preservation.

In Figure 10, we intimate just such a connectionist-described illusion, along the lines just discussed. Here, rather than activate associative loops based upon the content of the internal imagery, various associated feelings and impressions may nucleate as a result of the raw succession of internal imagery. That is, 'watching' and competing neural groupings may take notice of this parade of events and then act to attach varying degrees of significance to that succession. To some groupings it may appear that something lifelike (i.e., moving and evolving) resides within, providing a homuncular interpretation to the stream. Alternately, other subjective feelings may attach to the stream, including a qaule of awe over this ostensibly magical progression of intelligible information appearing from out of nowhere.

Peering into the cortex with various advanced technological tools such as PET and MRI, the changing pattern of activations within this and actual cortical layers consists of erratically fluctuating fields of 1's and 0's (i.e., neuron excitations and inhibitions, respectively), that in many respects resemble a turbulent binary sea (Figure 11). Internally, neural clusters may associate such digital pandemonium with similar chaotic phenomena in the external world such as in combustion and fluidic motion. This observation may help to explain the frequent literary allegory to the 'fires' and the 'seas' within, as well as our fascination with these phenomena. Taking this metaphor one step further, we may think of the myriad associative centers of the brain as predators, ever vigilant for internal imagery to emerge from the 'jungle grass'. As significant conceptual prey leaps from the clutter, feeding frenzy erupts, ending in the digestion of every last morsel by 'snarling' and 'warring' associative cascades.

---

that are commensurate with the loop (i.e., the loop contains an integer number of wavelengths) will survive by superposition. The recruited neurons will then appear to be in phase lock.

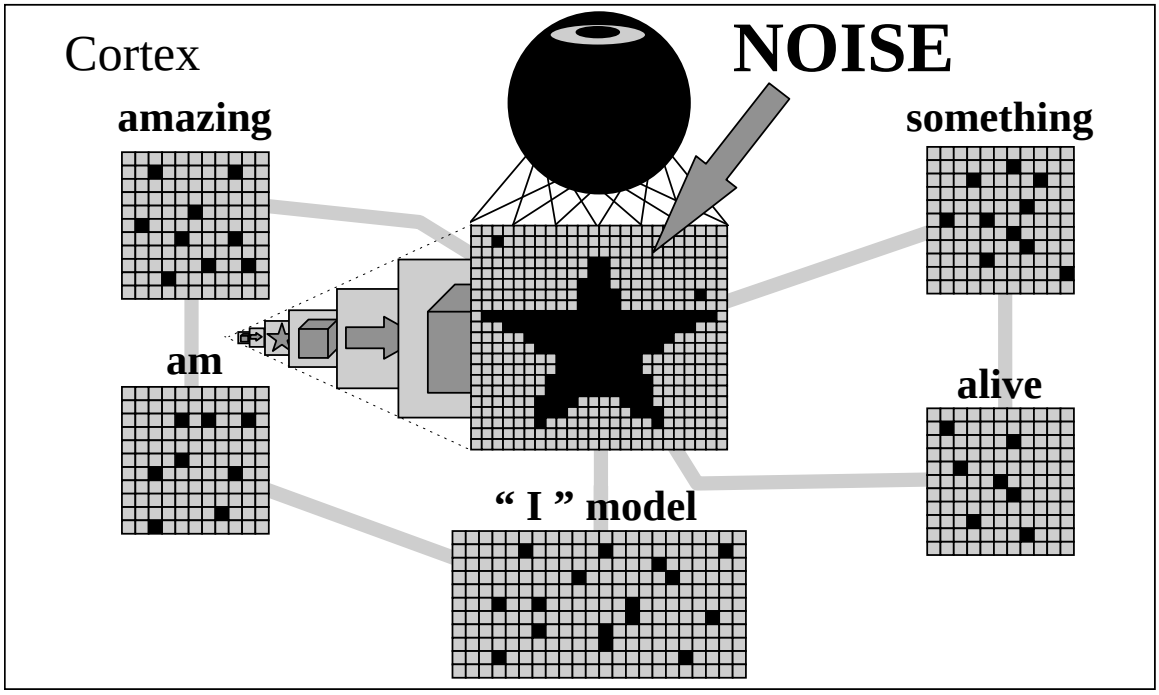


Figure 10. The raw stream of internal imagery within the simplified cortex nucleates a series of associated thoughts and impressions.

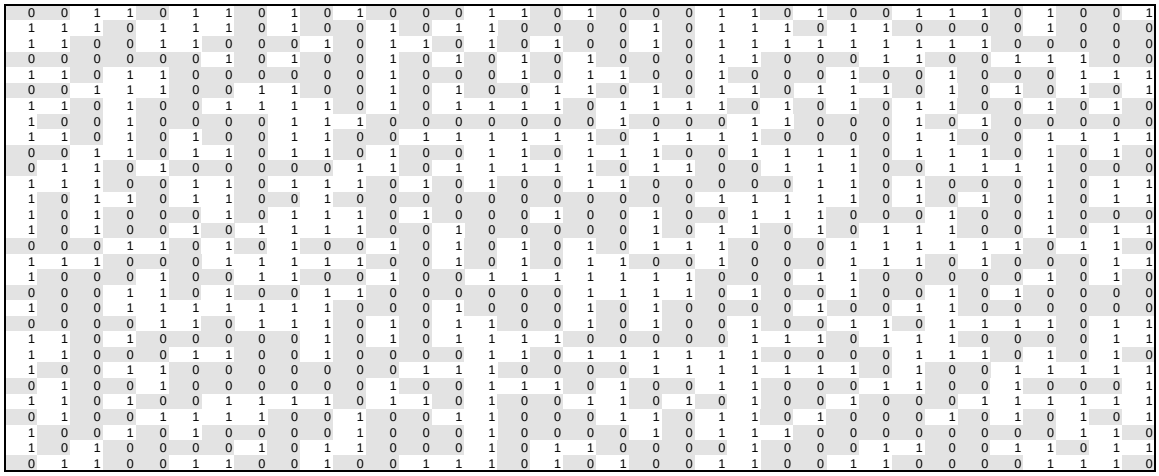


Figure 11. Digital chaos within a small section of cortex activates associative neural clusters related to notions of fire and sea.

I emphasize and underscore that these and other interpretations of global cortical chaos represent the gross approximations perpetrated by these lurking neural colonies. Such networks are performing reflex level interpretations of this inner activity so that what seems to be fire is not combustive at all. Again, such an impression is exceedingly illusory, along with most mental processing.

The theoretician’s associative loops

Within the cortex of an academician, this illusory process continues, as increasingly diverse neural clusters, relating to more abstruce concepts, compete to attach meaning to the overall stream of internal imagery. Of course, within each theoretician, there reside a number of neurologically habituated concepts that stand a

much better chance of successfully competing to assert their place within their theory (Figure 12). Therefore, those with a strong grounding in quantum mechanics (and its associated mysticism) are very likely to incorporate a wave functional interpretation to global cortical activity. Those immersed within early neurological reinforcement of spiritual concepts insist upon a non-material, non-reductionist dimension to any picture of consciousness. The reductionist, with a heavily habituated sense of minimalist explanation, applies the latest in the existing palette of physical principles, into a seemingly causal portrait. The point is that any combination of strongly 'burnt-in' paradigms dominate, integrating neurologically stored concepts of self (i.e., pride and personal investment) as well as overt or subconscious loyalties to various religious, political, or academic factions. In these very humble stochastic beginnings theories of consciousness arise. In retrospect, we glorify their origin as profound through a process related to Dennett's (1991) multiple drafts theory. Cults, academies, and the general population then adapt these notions within societal networks that in many respects resemble and function like neural networks.

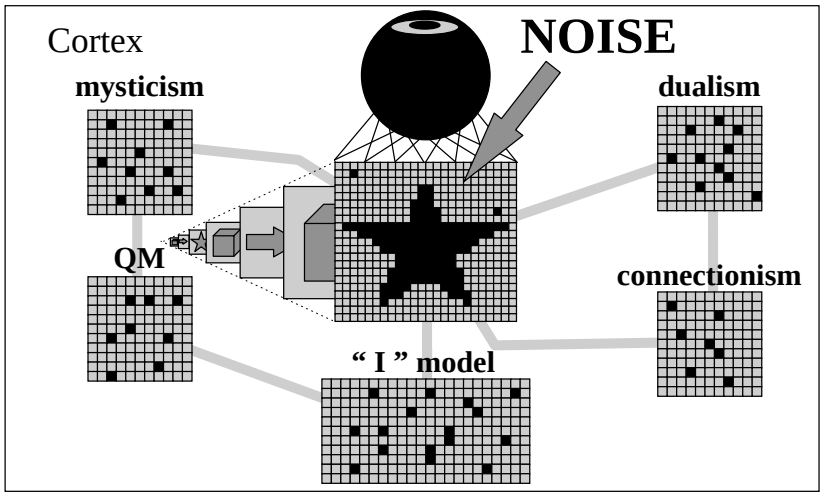


Figure 12. The raw stream of internal imagery within the simplified cortex of a theoretician nucleates a series of interpretations based upon the level of habituation of various neural groupings constituting the individual's analogy base.

The invention of meaning to the invention of meaning to the stream

Of course, everyone is conscious, so that all have some intuitive notion of what consciousness is. In connectionist's terms, this process corresponds to the activation of well-habituated associative networks, varying between individuals and representing alternative metaphors. These nets embody various automatic feelings about this purportedly profound and mysterious stream of thoughts and sensations. In other words, we all 'invent' significance to the stream. In turn, the societal network then attaches significance to our interpretations, lending weight and support to those it deems 'reasonable,' 'useful,' or politically advantageous. As in many scientific endeavors, evaluating institutions exercise ex cathedra priviledge in supporting those viewpoints on consciousness that fit their preconceptions, while muting those perspectives that don't.

VI: The Noise-Induced Stream and the Resulting Spectrum of Consciousness

Taking the perspective, that both the underlying stream of consciousness, as well as the interpretation of that stream are a function of internal noise within parallel-distributed processing systems, we may describe all levels of consciousness as the result of varying degrees of neurological perturbation (Figure 13). For instance, at internal noise levels corresponding to low level perturbations (i.e., leakage of neurotransmitters from presynaptic vesicles, cell membrane potential fluctuations, crosstalk, etc.) there are sufficiently low

levels of network disruption so as to activate largely intact memories. Such a stream of straightforward environmental images and their interpretation comprise normal waking stream of consciousness.

At slightly higher levels of internal perturbation, where for example long term potentiation within synapses may relax, as in REM sleep, such localized disturbances may propagate throughout the cortex to activate a series of memories both straightforward as well as corrupted. The result is dreaming, something all parallel processing systems tend do when cut off from external sensory inputs and exposed to internal chaos. Perpetually alert policing networks within the cortex may activate in response to these internal 'rehashes' of the environment, should any emerging image bear survival value. Should this review contain the imagery of real or confabulated threats in the external world, the result is a nightmare, providing an invaluable rehearsal for similar waking encounters.

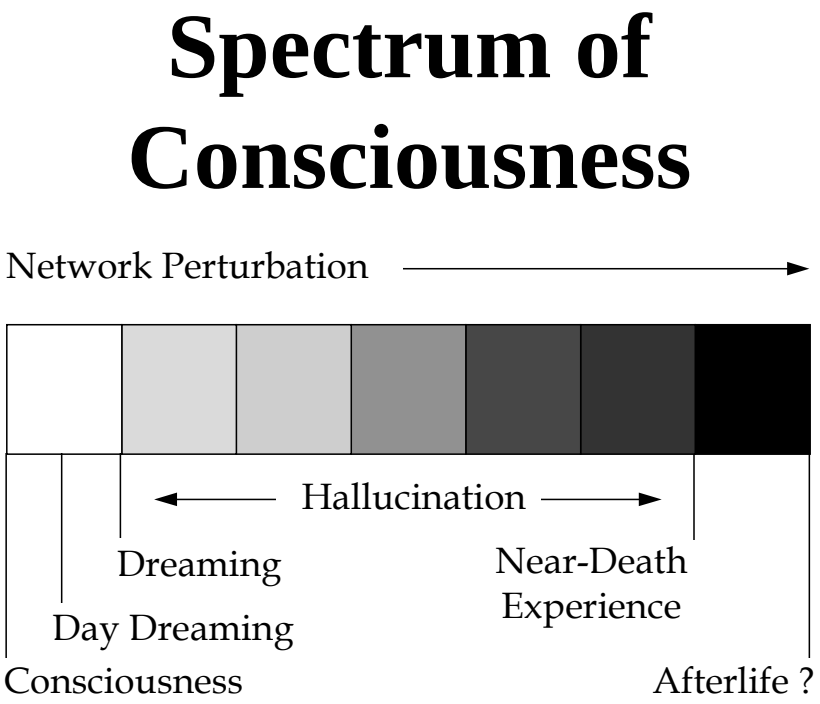


Figure 13. The proposed spectrum of consciousness is largely a function of the degree of internal neurological perturbation. (S.L. Thaler, 1996d)

As we further increase the level of internal disruption synaptic meddling, either with internally produced neurotransmitters or their imitations (i.e., drugs), the network mapping degrades sufficiently to produce a succession of largely confabulatory experience that we call hallucination. We may experience such virtual inner worlds through the trauma induced secretion of neuromodulators and neurohormones.

Finally at the most extreme levels of perturbation possible within connectionist systems, metabolic death annihilates whole processing units and their synaptic links. The quickly condensing cortical networks now produce abundant damage now interpreted as a torrent of highly confabulated virtual experience by the remaining intact portions of the network cascade. At its onset, we duplicate the features of life review and later as damage rapidly avalanches, a torrent of novel events ensues, tantamount to the much mythologized near-death experience.

In effect then, most mental experience is an internally generated illusion, sporadically peppered with highly degraded news of incidents in the external world. We may therefore view all of this experience as various degrees of dreaming, mediated by multiple sources of internal chaos.

VII. Summary

I have proposed a pragmatic theory of consciousness that presupposes absolutely no special significance of human kind in the overall functioning or destiny of the universe. It consists of three broad assertions that attach more of a devolutionary rather than evolutionary significance to this phenomenon.

1. That the dissolution of interactions (i.e., connections) within portions of the universe has led to the creation of insular parallel processing[2] regions whose relatively strong inner connectivity leads to an accompanying numbness or insensitivity to the external universe. Weak couplings (i.e., the senses) with that outer world allow for sporadic communication between the external environment and these isolated regions in a process tantamount to perception.

2. Plentiful chaos within these insular regions tends to drive these systems through a series of activation states representative of conditions in the external environment (i.e., internal imagery). Further internal fragmentation of these regions produce connectionist sub-modules that may react to scenarios imagined within other sub-modules, 'seizing' upon imagined concepts of utility, interest or survival value (i.e., the Creativity Machine Paradigm).

3. Only those isolated connectionist clusters that have per chance developed an 'illusory' network mapping relating overall perception and internal imagery to a quale of being, distinctness, and self-importance will self-organize to survive, sprouting the necessary manipulative, locomotive and reproductive paraphernalia. Among these persisting isolationist clusters are the human organisms. Otherwise, such insular connectionist regions simply reconnect to the whole, lost to the notion of a separate identity.

We, as human beings, are recent instantiations of these connectionist islands, in possession of an illusory neural network mapping that funnels the raw succession of chaos-driven internal imagery and rapidly converts it to such 'super-qualia' as 'self' and of 'being' (as depicted in Figure 14).[3] These feelings are no more than the outcome of cognitive combat between neural factions to attach meaning to overall cortical activity. The winner in this battle is the subjective experience of consciousness or the 'buzz,' if you will.

Therefore, if asked to supply a summary definition of consciousness, I would feel compelled to clearly and adamantly state the following -

Consciousness is the involuntary invention of significance to overall brain activity.

Implicit in this definition is the notion that this significance consists of a family of super-qualia that foster feelings of separateness, self worth and self-preservation.

It may be that these super-qualia (an inescapable neurological feature of all of us who study consciousness) ultimately stand in the way of an accurate model of consciousness. These feelings subtly and inexorably protest at the mechanical models, constantly reminding us of the richness and diversity of subjective experience. They deceive us into believing that conscious experience is deeply profound, requiring equally soaring explanations.

However most leading connectionists are willing to commit to battle against this natural deception. To paraphrase the renown neurophilosopher P.S. Churchland, (1996), a comprehensive model of consciousness will not require a "humdinger" explanation. Certainly, within the perspective I have adopted, we require no new physics, only the open mindedness to admit that our most intuitive

[2] That is, all interactions among entities occur simultaneously as if all of these items are independent microprocessors contemporaneously evaluating the effect of all others upon it.

[3] Liberation from such super-qualia of self and of being has been reported to occur during temporal lobe seizure and neurosurgical procedures (Morse et al., 1989 and Penfield and Rasmussen, 1950). These results tend to anatomically pinpoint the seat of this illusory network mapping. It is the destruction of this super-quale of self that tends to produce the anguish within negative near-death experience (Ring, 1994).

preconception of consciousness is a hard-wired and self-protective neurobiological illusion. Out of Churchland's brand of synaptic plasticity arises a model that allows for the continuity of human consciousness with other concurrent physical processes within the universe.

Again, I emphasize that what I and other connectionists are proposing is not the ultimate theory, but a more pragmatic model that genuinely supports the objectives of prediction, technological application, and yes, even psychological comfort. In the first regard, we may statistically define stream of consciousness by Equation 2, with the content of cognition contained within the details of neuronal activation. In the second matter of technological utility, the Creativity Machine paradigm is yielding unanticipated human-level discovery across all fields of human endeavor.[4] In the final psychological matter, the model may provide a profound and real alternative to the 'other world' succor against the fear of death once marketed by Becker (1973). Its main features of this new psychological haven are as follows:

1. That non-biological connectionist islands may also be capable of consciousness, thus generalizing this once sacred quality to strictly inorganic, physical systems. Thus Arthur Clarke's and originally Olaf Stapledon's view of a cosmic consciousness (Ash, 1977) attains even higher significance and reality. The mind-boggling possibility is that we may ultimately be capable of reconnecting, through massively parallel connections, to various so-called inanimate objects such as trees, rocks, and stars, and directly feeling their consciousness. (Do not raise your expectations though, because all the connectionist activations within these objects can only be interpreted by the human brain as human experience. Thus the thoughts of a star can only be projected into the realm of human concepts and feelings.)

2. That we may personally experience the consciousness of other biological and non-biological entities by re-establishment of connections between them and ourselves. This achievement will require a technology that supplies not just a few isolated channels of communication between agencies, but a massively parallel one, as alluded to in principle 1. Once intuitively in communication with another biological sentience, consciousness may join like two fusing drops of water. We may thus mind-meld with specially prepared clones, sharing cumulative cognitive and emotional experience. As the old body withers and dies, it will be much the same as an amputated limb. Biological immortality will thereby be achieved.

3. That we may merge with non-biological entities of superior robustness to the human biological housing. To this end, we may assemble computational vehicles or 'shells', allowing us to freely wander and investigate the universe, or to think profound, unthought of thoughts, ad infinitum.

4. That as many nonreductionist personalities have claimed, death is not final, but a transitory experience. In the technical model proposed herein, this transition is a process of reconnection with the universe as the specialized networks responsible for the illusionary quale of self and distinctness dissolve.

[4] Of course there will be an immense sociological battle to drive home this significant accomplishment.

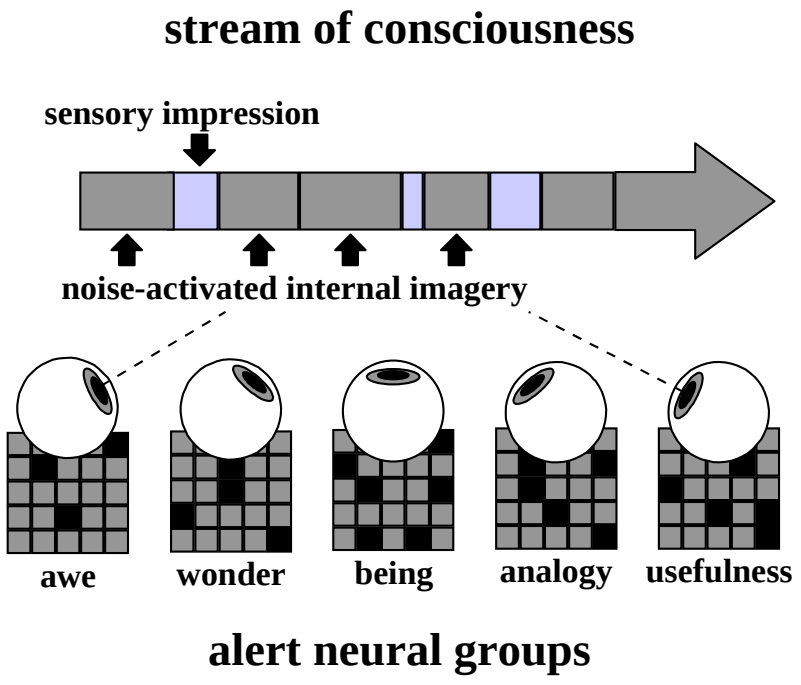


Figure 14. Stream of consciousness, driven by internal noise, is interpreted by observing networks as various 'super-qualia' or illusions normally associated with consciousness.

It is to be noted that in the context of the two latter principles, that there will always be a 'rush' of virtual and novel experience accompanying any significant amount of connective modification. In one instance this experiential surge will come from the addition of connectionist units and in the latter from the elimination of such units. In either case, there will be the equivalent of near-death experience, as vector completion goes to work interpreting neural damage, annexation, or both.

In concluding and extrapolating, the choice we now face is whether to cart the illusory part of us along for the ride or leave it to decay in the soil of earth. Those who pack lightly will fade into connectivity with all else. Those who retain this cargo will, by definition, remain distinct.

References

Ash, B. (ed.), *The Visual Encyclopedia of Science Fiction* (Copplestone Publishing Ltd.), p 201.

Becker, E. (1973). *The Denial of Death* (Free Press)

Chalmers, D.J. (1996). On the search for the neural correlate of consciousness, *Consciousness Research Abstracts, Toward a Science of Consciousness 1996 "Tuscon II"*, abstract 98, p. 61.

Churchland, P.M (1996), Our friend the microtubule, *Consciousness Research Abstracts, Toward a Science of Consciousness 1996 "Tuscon II"*, abstract 143, p. 74.

Dennett, D.C. (1991). *Consciousness Explained* (Boston: Little, Brown, and Co)

Edelman, G.M. (1989). The Remembered Present: A Biological Theory of Consciousness (New York: Basic Books)

Franklin, S. (1995). *Artificial minds* (Boston:MIT Press)

Freeman, W.J., and Skarda, C.S. (1990). "Representations: Who Needs Them?" J.L. McGaugh, et al., eds., *Brain Organization and Memory Cells, Systems and Circuits* (New York: Oxford University Press)

Gleick, J. (1987), *Chaos, making a new science* (New York: Penguin Group), p8.

Morse, M.L., Venecia, D. & Milstein, J. (1989). Near-Death Experiences: A Neurophysiologic Explanatory Model, *Journal of Near-Death Studies*, **8**(1) , pp. 45-53.

Penfield, W. & Rasmussen, T. (1950). The cerebral cortex of man: A clinical localization of function (New York: Macmillan)

Ring, K. (1994). Solving the riddle of frightening near-death experiences: some testable hypotheses and a perspective based on a course in miracles, *Journal of Near-Death Studies*, **13**(1), pp. 5-23.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition.* **Vol. 1** (eds Rumelhart, D.E. & McClelland, J.L.) (Boston: MIT Press)

Thaler, S.L. (1995). "Virtual input" phenomena within the death of a simple pattern associator, *Neural Networks*, **8**(1), pp. 55-66.

Thaler, S.L. (1996c). 'Network cavitation' in the modeling of consciousness, *Consciousness Research Abstracts, Toward a Science of Consciousness 1996 "Tuscon II"*, abstract 230, p. 102.

Thaler, S.L. (1996b). Is neuronal chaos the source of stream of consciousness?, to be published in Proceedings of the World Congress on Neural Networks, WCNN'96, Lawrence Erlbaum & Associates.

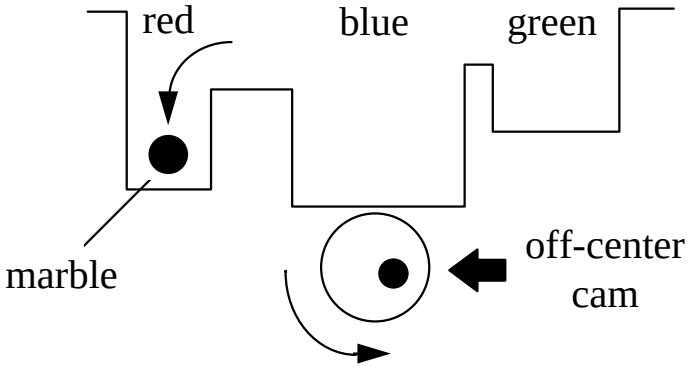
Thaler, S.L. (1996a). Neural networks that create and discover, *PCAI*, May/June 1996, pp. 16-21.

Thaler, S.L. (1996d). The death dream and near-death Darwinism, *Journal of Near-Death Studies*, **15**(1) , pp. 25-40.

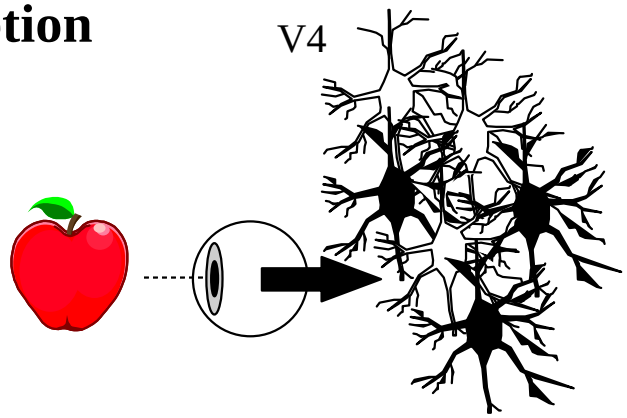
Voss, R. (1988) in *The Science of Fractal Images* (eds Peitgen, H. & Saupe, D.), 21-70, Springer-Verlag.

SIDE BAR

attractor basins

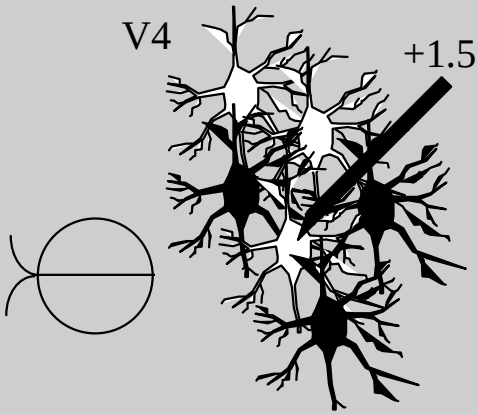


perception



red quale activated by sensory input

internal imagery



red quale activated by surgeon's probe or internal chaos

Side Bar Text

Connectionists generally regard any memory or idea as a specific on / off pattern of cortical neurons. Generally speaking, these ***activation patterns*** are distributed over large areas of cortex rather than being localized to specific cells or, in the other extreme, totally distributed in a 'holographic' manner. Using a group of neurons to represent an oversimplified color vision center (corresponding to the V4 area of visual cortex) the perception of redness from an apple takes the form of an activation pattern within that colony of cells. Here we depict 'on' neurons as white and 'off' as black. This is an important notion to connectionists in that not only can color sensation be represented by such on / off configurations of brain cells, but also concepts, sensations, emotions, and other general qualia. That these feeling seem so complex is that they launch activations in myriad connected and associated neural colonies. The original perception or imagery is then nested within a complex pattern of cell firings widely distributed over cortex.

The notion of an ***attractor basin*** is illustrated at the top of the side bar by a three-pocketed tray vibrated by a rapidly rotating cam. A marble is thereby agitated to hop between these pockets, generally corresponding to either red, blue, or green qualia being randomly activated in our pedagogical color cortex. Rather than being geometrically constrained in the vibrating tray and gravitation, the V4 center is confined to visit only three distinct states established by the balance of excitatory and inhibitory connections between neurons. If a red signal, specified by a vector or pattern of activations by retinal cones matches the on / off pattern of redness in cortex, the quale of redness and all its associations are launched.

Alternately, the identical quale of red may be stimulated accidentally by a surgeon probing V4 with an electrified probe. The colony of neurons, V4, performs what is called ***vector completion*** on this spurious signal, activating into an overall pattern that best matches the pattern of perturbation introduced by the probe. We all do this in other contexts and sensory modalities when we unconsciously complete a sentence, overlook typographical errors, or imagine a cloud to resemble an animal. Similarly, internal perturbation similar in effect to the electrode, may stimulate a counterfeit, yet indistinguishable feeling of redness. This effect is what I have observed in artificial neural networks and coined ***virtual input effect*** to convey the perception of red without the physical presence of red, due to myriad forms of internal network perturbations.

Germane to the main theme of this article, both the probe, perhaps a stainless steel or glass micropipette, and the cortex are ***connectionist systems***. In the former, the connections amount to the myriad electrical and magnetic interactions that bind atoms together as well as the needle's continuity with a voltage supply. In the latter case, in addition to such atomistic considerations, neurons are likewise bound together into a collective unit. Both have 'feelings' in that any portion of their mass may react to activity in any other portion. For instance for the needle, mechanical stress (or electrical potential) at one end of its length is sensed at the opposite end as atoms there minutely reposition themselves in response. In the cortex, electrical activation of any cell or group of cells will result in a distributed reactivation of many other cells. When these two connectionist systems meet, they are effectively interconnected, yet the cortex cannot sense what it feels like to be a probe. Instead, the cortex will activate to feel any of its colors, while the needle will continue to feel its many internal physical states. Not until such systems are massively interconnected will there be a mutual 'understanding' of each others' stream of internal impressions.