CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Nature Precedings

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

Version 7

When Spandrels Become Arches: Neural crosstalk and the evolution of consciousness

Rodrick Wallace
Division of Epidemiology
The New York State Psychiatric Institute *

July 14, 2011

Abstract

Once cognition is recognized as having a 'dual' information source, the information theory chain rule implies that isolating coresident information sources from crosstalk requires more metabolic free energy than permitting correlation. This provides conditions for an evolutionary exaptation leading to the rapid, shifting global neural broadcasts of consciousness. The argument is quite analogous to the well-studied exaptation of noise to trigger stochastic resonance amplification in neurons and neuronal subsystems. Astrobiological implications are obvious.

Key Words: astrobiology, information theory, phase transition, sufficient conditions

1 Introduction

Researchers have long speculated and experimented on the role of noise in neural processes and subsystems via models of stochastic resonance (e.g., Park and Neelakanta, 1996; Gluckman et al., 1996; Ward, 2009; Kawaguchi et al., 2011). The necessary ubiquity of noise affecting information transmission underwent an evolutionary exaptation (e.g., Gould, 2002) to become a tool for amplification of weak signals. Here we examine the parallel necessary circumstance of information leakage between 'adjacent' communication channels or information sources, a generally unwelcome signal correlation that the electrical engineers

*Box 47, 1051 Riverside Dr., New York, NY, 10032, rodrick.wallace@gmail.com

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

call 'crosstalk'. The evolutionary exaptation of crosstalk appears to be the system of rapid, shifting global neural broadcasts we characterize as consciousness.

Baars' global workspace model of animal consciousness attributes the phenomenon to a shifting array of unconscious cognitive modules that unite to become a global broadcast having a tunable perception threshold not unlike a theater spotlight whose range of attention is constrained by embedding contexts (e.g., Baars, 1988, 2005; Baars and Franklin, 2003). The basic mechanism emerges 'naturally' from a remarkably simple application of the asymptotic limit theorems of information theory, once a broad range of unconscious cognitive processes is recognized as inherently characterized by information sources – generalized languages (Wallace, 2000, 2005, 2007). The approach allows mapping physiological unconscious cognitive modules onto an abstract network of interacting information sources. This, in turn, permits a simplified mathematical attack based on phase transitions in network topology that, in the presence of sufficient linkage – crosstalk – permits rapid, shifting, global broadcasts. While the mathematical description of consciousness is itself relatively simple, the evolutionary trajectories leading to its emergence seem otherwise. Here we argue that this is not the case, and that physical restrictions on the availability of metabolic free energy provide sufficient conditions for the emergence of consciousness.

The argument is, in a sense, an extension of Gould and Lewontin's (1979) famous essay "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme". Spandrels are the triangular sectors of the intersecting arches that support a cathedral roof. They are simple byproducts of the need for arches, and their occurrence is in no way fundamental to the construction of a cathedral. Our assertion is that crosstalk between unconscious cognitive modules is a similar inessential byproduct that evolutionary process has exapted to construct the rapidly shifting global broadcasts of consciousness: Evolution built a new arch from a spandrel.

We first provide a minimal formal overview that will be reexpressed in more complex form, much like Onsager's nonequilibrium thermodynamics.

2 Cognition as 'language'

Atlan and Cohen (1998) argue, in the context of a cognitive paradigm for the immune system, that the essence of cognitive function involves comparison of a perceived signal with an internal, learned or inherited picture of the world, and then, upon that comparison, choice of one response from a much larger repertoire of possible responses. That is, cognitive pattern recognition-and-response proceeds by an algorithmic combination of an incoming external sensory signal with an internal ongoing activity – incorporating the internalized picture of the world – and triggering an appropriate action based on a decision that the pattern of sensory activity requires a response.

More formally, incoming sensory input is mixed in an unspecified but systematic algorithmic manner with a pattern of internal ongoing activity to create

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

a path of combined signals $x = (a_0, a_1, ..., a_n, ...)$. Each a_k thus represents some functional composition of the internal and the external. An application of this perspective to a standard neural network is given in Wallace (2005, p.34).

This path is fed into a highly nonlinear, but otherwise similarly unspecified, decision oscillator, h, which generates an output $h(x)$ that is an element of one of two disjoint sets B_0 and B_1 of possible system responses. Let

$$B_0 \equiv \{b_0, ..., b_k\},$$

$$B_1 \equiv \{b_{k+1}, ..., b_m\}.$$

Assume a graded response, supposing that if

$$h(x) \in B_0,$$

the pattern is not recognized, and if

$$h(x) \in B_1,$$

the pattern is recognized, and some action $b_j, k+1 \leq j \leq m$ takes place.

The principal objects of formal interest are paths x which trigger pattern recognition-and-response. That is, given a fixed initial state a_0, we examine all possible subsequent paths x beginning with a_0 and leading to the event $h(x) \in B_1$. Thus $h(a_0, ..., a_j) \in B_0$ for all $0 < j < m$, but $h(a_0, ..., a_m) \in B_1$.

For each positive integer n, let $N(n)$ be the number of high probability grammatical and syntactical paths of length n which begin with some particular a_0 and lead to the condition $h(x) \in B_1$. Call such paths 'meaningful', assuming, not unreasonably, that $N(n)$ will be considerably less than the number of all possible paths of length n leading from a_0 to the condition $h(x) \in B_1$.

While combining algorithm, the form of the nonlinear oscillator, and the details of grammar and syntax, are all unspecified in this model, the critical assumption which permits inference on necessary conditions constrained by the asymptotic limit theorems of information theory is that the finite limit

$$H \equiv \lim_{n \to \infty} \frac{\log[N(n)]}{n} \tag{1}$$

both exists and is independent of the path x.

Call such a pattern recognition-and-response cognitive process *ergodic*. Not all cognitive processes are likely to be ergodic, implying that H, if it indeed

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

exists at all, is path dependent, although extension to nearly ergodic processes, in a certain sense, seems possible (e.g., Wallace, 2005, pp. 31-32).

Invoking the spirit of the Shannon-McMillan Theorem, it is possible to define an adiabatically, piecewise stationary, ergodic information source $\mathbf{X}$ associated with stochastic variates X_j having joint and conditional probabilities $P(a_0, ..., a_n)$ and $P(a_n|a_0, ..., a_{n-1})$ such that appropriate joint and conditional Shannon uncertainties satisfy the classic relations

$$H[\mathbf{X}] = \lim_{n\to\infty} \frac{\log[N(n)]}{n} =$$

$$\lim_{n\to\infty} H(X_n|X_0, ..., X_{n-1}) =$$

$$\lim_{n\to\infty} \frac{H(X_0, ..., X_n)}{n}.$$

(2)

This information source is defined as *dual* to the underlying ergodic cognitive process, in the sense of Wallace (2000, 2005).

The essence of 'adiabatic' is that, when the information source is parameterized according to some appropriate scheme, within continuous 'pieces' of that parameterization, changes in parameter values take place slowly enough so that the information source remains as close to stationary and ergodic as needed to make the fundamental limit theorems work. By 'stationary' we mean that probabilities do not change in time, and by 'ergodic' (roughly) that cross-sectional means converge to long-time averages. Between 'pieces' one invokes various kinds of phase change formalism, for example renormalization theory in cases where a mean field approximation holds (Wallace, 2005), or variants of random network theory where a mean number approximation is applied. More will be said of this latter approach below.

Recall that the Shannon uncertainties $H(...)$ are cross-sectional law-of-large-numbers sums of the form $-\sum_k P_k \log[P_k]$, where the P_k constitute a probability distribution. See Cover and Thomas (2006), Ash (1990), or Khinchin (1957) for the standard details.

3 Dynamic networks of unconscious cognitive modules and the 'no free lunch' theorem

The famous 'no free lunch' theorem of Wolpert and Macready (1995, 1997) illuminates the next step in the argument. As English (1996) states the matter,

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

> ...Wolpert and Macready... have established that there exists no generally superior [computational] function optimizer. There is no 'free lunch' in the sense that an optimizer 'pays' for superior performance on some functions with inferior performance on others... if the distribution of functions is uniform, then gains and losses balance precisely, and all optimizers have identical average performance... The formal demonstration depends primarily upon a theorem that describes how information is conserved in optimization. This Conservation Lemma states that when an optimizer evaluates points, the posterior joint distribution of values for those points is exactly the prior joint distribution. Put simply, observing the values of a randomly selected function does not change the distribution...
>
> [A]n optimizer has to 'pay' for its superiority on one subset of functions with inferiority on the complementary subset...
>
> Anyone slightly familiar with the [evolutionary computing] literature recognizes the paper template 'Algorithm X was treated with modification Y to obtain the best known results for problems P_1 and P_2.' Anyone who has tried to find subsequent reports on 'promising' algorithms knows that they are extremely rare. Why should this be?
>
> A claim that an algorithm is the very best for two functions is a claim that it is the very worst, on average, for all but two functions.... It is due to the diversity of the benchmark set [of test problems] that the 'promise' is rarely realized. Boosting performance for one subset of the problems usually detracts from performance for the complement...
>
> Hammers contain information about the distribution of nail-driving problems. Screwdrivers contain information about the distribution of screw-driving problems. Swiss army knives contain information about a broad distribution of survival problems. Swiss army knives do many jobs, but none particularly well. When the many jobs must be done under primitive conditions, Swiss army knives are ideal.
>
> The tool literally carries information about the task... optimizers are literally tools-an algorithm implemented by a computing device is a physical entity...

Another way of stating this conundrum is to say that a computed solution is simply the product of the information processing of a problem, and, by a very famous argument, information can never be gained simply by processing. Thus a problem X is transmitted as a message by an information processing channel, Y, a computing device, and recoded as an answer. By the extended argument of the Mathematical Appendix, there will be a channel coding of Y which, when properly tuned, is most efficiently 'transmitted', in a purely formal sense, by the problem. In general, then, the most efficient coding of the transmission channel, that is, the best algorithm turning a problem into a solution, will necessarily be highly problem-specific. Thus there can be no best algorithm for all sets of

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

problems, although there will likely be an optimal algorithm for any given set.

Based on the no free lunch argument of the previous section, it is clear that different challenges facing an entity must be met by different arrangements of basic unconscious cognitive modules. It is possible to make a very abstract picture of this phenomenon, not based on neural anatomy, but rather on the linkages between the information sources dual to the basic physiological and learned unconscious cognitive modules (UCM). That is, *the remapped network of unconscious cognitive modules is reexpressed in terms of the information sources dual to the UCM.* Given two distinct problems classes (e.g., the search for food vs. reproduction), there must be two different 'wirings' of the information sources dual to the physiological UCM, as in figure 1, with the network graph edges measured by the amount of information crosstalk between sets of nodes representing the dual information sources. A more formal treatment of such coupling can be given in terms of network information theory (Cover and Thomas, 2006), particularly incorporating the effects of embedding contexts, implied by the 'external' information source Z – signals from the environment.

The possible expansion of a closely linked set of information sources dual to the UCM into a global workspace/broadcast – the occurrence of a kind of 'spandrel' – depends, in this model, on the underlying network topology of the dual information sources and on the strength of the couplings between the individual components of that network. For random networks the results are well known, based on the work of Erdos and Renyi (1960). Following the review by Spenser (2010) closely (see, e.g., Boccaletti et al., 2006, for more detail), assume there are n network nodes and e edges connecting the nodes, distributed with uniform probability – no nonrandom clustering. Let $G[n, e]$ be the state when there are e edges. The central question is the typical behavior of $G[n, e]$ as e changes from 0 to $(n-2)!/2$. The latter expression is the number of possible pair contacts in a population having n individuals. Another way to say this is to let $G(n, p)$ be the probability space over graphs on n vertices where each pair is adjacent with independent probability p. The behaviors of $G[n, e]$ and $G(n, p)$ where $e = p(n-2)!/2$ are asymptotically the same.

For 'real world' biological and social structures, one can have $p = f(e, n)$, where f may not be simple or even monotonic. For example, while low e would almost always be associated with low p, beyond some threshold, high e might drive individuals or nodal groups into isolation, decreasing p and producing an 'inverted-U' signal transduction relation akin to stochastic resonance. Something like this would account for Fechner's law which states that perception of sensory signals often scales as the log of the signal intensity.

For the simple random case, however, we can parameterize as $p = c/n$. The graph with $n/2$ edges then corresponds to $c = 1$. The essential finding is that the behavior of the random network has three sections:

[1] If $c < 1$ all the linked subnetworks are very small, *and no global broadcast can take place.* This is taken as the standard operating mode for nonminded organisms.

[2] If $c = 1$ there is a single large interlinked component of a size $\approx n^{2/3}$.

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

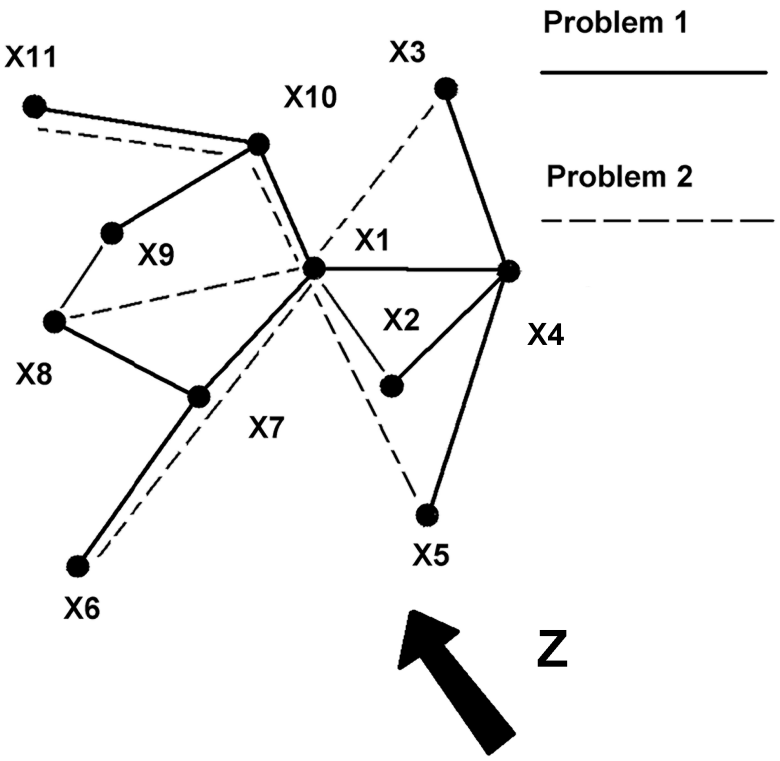

Figure 1: By the no free lunch theorem, two markedly different problems will be optimally solved by two different linkages of available unconscious cognitive modules – characterized now by their dual information sources X_j – into different temporary networks of working structures, here represented by crosstalk among those sources rather than by the physiological UCM themselves. The embedding information source Z represents the influence of external signals whose effects can be accounted for by an application of network information theory.

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

[3] If $c > 1$ then there is a single large component of size yn – a global broadcast – where y is the positive solution to the equation

$$\exp(-cy) = 1 - y. \tag{3}$$

Then

$$y = \frac{W(-c/\exp(c)) + c}{c}, \tag{4}$$

where W is the Lambert W function.

The solid line in figure 2 shows y as a function of c, representing the fraction of network nodes that are incorporated into the interlinked giant component – a de-facto global broadcast for interacting UCM. To the left of $c = 1$ there is no giant component, and large scale cognitive process is not possible.

The dotted line, however, represents the fraction of nodes in the giant component for a highly nonrandom network, a star-of-stars-of-stars (SoS) in which every node is directly or indirectly connected with every other one. For such a topology there is no threshold, only a single giant component, showing that the emergence of a giant component in a network of information sources dual to the UCM is dependent on a network topology that may itself be tunable.

4 Information and free energy: how a spandrel can become an arch

The information sources dual to unconscious cognitive modules represented in figure 1 are not independent, but are correlated, so that a joint information source can be defined having the properties

$$H(X_1, ..., X_n) \leq \sum_{j=1}^{n} H(X_j).$$

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

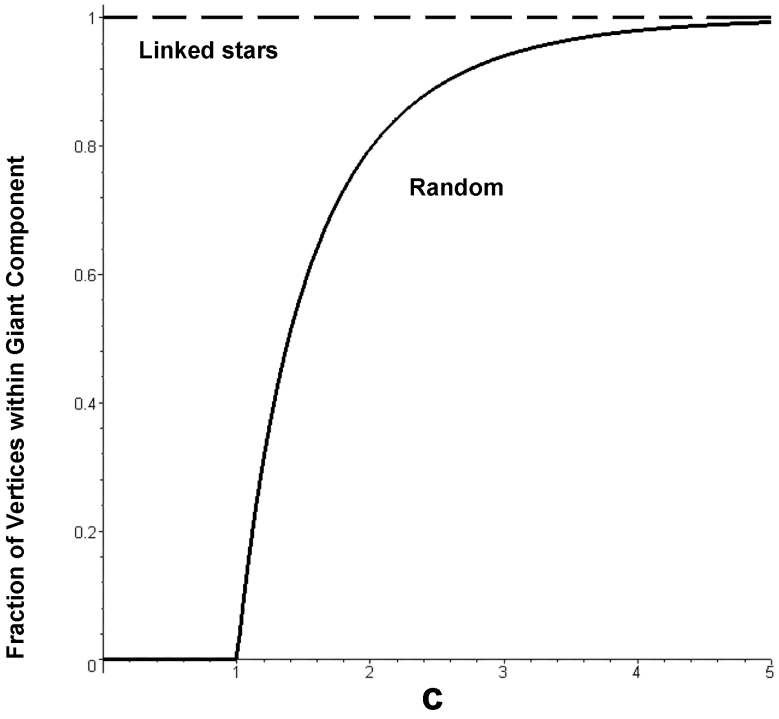

Figure 2: Fraction of network nodes in the giant component as a function of the crosstalk coupling parameter c. The solid line represents a random graph, the dotted line a star-of-stars-of-stars network in which all nodes are interconnected, showing that the dynamics of giant component emergence are highly dependent on an underlying network topology that, for UCM, may itself be tunable. For the random graph, a strength of $c < 1$ precludes emergence of a global broadcast.

(5)

This result is known as the information chain rule (e.g., Cover and Thomas, 2006), and has a very profound implication. As Feynman (2000) describes at length, information is, in fact, a form of free energy. Feynman even shows how to construct a simple (ideal) machine that can convert the information in a message into actual work, the essence of a free energy argument. By the chain rule, then, it takes more metabolic free energy to keep cognitive modules from interacting than it does to let them become correlated via crosstalk. As the electrical engineers tell us, preventing unwanted crosstalk takes a considerable investment of resources.

The global broadcast mechanisms of consciousness make an arch of this spandrel, using the lowered free energy requirement of crosstalk interaction between unconscious cognitive modlules as the springboard for launching the rapid, tunable, highly correlated, global broadcasts that we characterize as consciousness.

5 Reaction to environmental signals

Unconscious cognitive modules operate within larger, highly structured, environmental signals and other constraints whose regularities may also have a recognizable grammar and syntax, represented in figure 1 by an embedding information source Z. Under such a circumstance the splitting criterion for three jointly typical sequences is given by the classic relation of network information theory (Cover and Thomas, 2006, Theorem 15.2.3)

$$I(X_1, X_2|Z) = H(Z) + H(X_1|Z) + H(X_2|Z) - H(X_1, X_2, Z)$$

(6)

that generalizes as

$$I(X_1, ..., X_n|Z) = H(Z) + \sum_{j=1}^{n} H(X_j|Z) - H(X_1, ..., X_n, Z)$$

(7)

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

More complicated multivariate typical sequences are treated much the same (e.g., El Gamal and Kim, 2010, p.2-26). Given a basic set of interacting information sources $(X_1, ..., X_k)$ that one partitions into two ordered sets $X(\mathcal{J})$ and $X(\mathcal{J}')$, then the splitting criterion becomes $H[X(\mathcal{J}|\mathcal{J}')]$. Extension to a greater number of ordered sets is straightforward.

Then the joint splitting criterion – I, H above – however it may be expressed as a composite of the underlying information sources and their interactions, satisfies a relation like the first expression in equation (2), where $N(n)$ is the number of high probability jointly typical paths of length n, and the theory carries through, now incorporating the effects of external signals as the information source Z.

6 A formal model

Given the splitting criteria $I(X_1, ..., X_n|Z)$ or $H[X(\mathcal{J}|\mathcal{J}')]$ as above, the essential point is that these are the limit, for large n, of the expression $\log[N(n)]/n$, where $N(n)$ is the number of jointly typical paths of the interacting information sources of length n. Again, as Feynman (2000) argues at great length, information is simply another form of free energy, and its dynamics can be expressed using a formalism similar to Onsager's nonequilbrium thermodynamics.

The argument is direct.

First, the physical model. Let $F(K)$ be the free energy density of a physical system, K the normalized temperature, V the volume and $Z(K, V)$ the *partition function* defined from the Hamiltonian characterizing energy states E_i. Then

$$Z(V, K) \equiv \sum_i \exp[-E_i(V)/K], \tag{8}$$

and

$$F(K) = \lim_{V \to \infty} -K \frac{\log[Z(V, K))}{V} \equiv \frac{\log[\hat{Z}(K, V)]}{V}, \tag{9}$$

If a nonequilibrium physical system is parameterized by a set of variables $\{Q_i\}$, then the *empirical Onsager equations* are defined in terms of the gradient of the entropy $S \equiv F - \sum_j Q_j dF/dQ_j$ as

$$dQ_j/dt = \sum_i L_{i,j} \partial S/\partial Q_i,$$

(10)

where the $L_{i,j}$ are empirical constants. For a physical system having microreversibility, $L_{i,j} = L_{j,i}$. For an information source where, for example, ' the ' has a much different probability than ' eth ', no such microreversibility is possible, and no 'reciprocity relations' can apply.

For stochastic systems this generalizes to the set of stochastic differential equations

$$dQ_t^j = \sum_i [L_{j,i}(t, ...\partial S/\partial Q^i ...)dt + \sigma_{j,i}(t, ...\partial S/\partial Q^i)dB_t^i]$$

$$= L(Q^1, ..., Q^n)dt + \sum_i \sigma(t, Q^1, ..., Q^n)dB_t^i,$$

(11)

where terms have been collected and expressed in terms of the driving parameters. The dB_t^i represent different kinds of 'noise' whose characteristics are usually expressed in terms of their quadratic variation. See any standard text for definitions, examples, and details.

The essential trick is to recognize that, for the splitting criteria $I(X_1, ..., X_n|Z)$ or $H[X(\mathcal{J}|\mathcal{J}')]$, the role of information as a form of free energy, and the corresponding limit in $\log[N(n)]/n$, make it possible to define entropy-analogs as

$$S \equiv I(...Q^k...) - \sum_j Q^j \partial I/\partial Q^j$$

$$S \equiv H[X(\mathcal{J}|\mathcal{J}')] - \sum_j Q^j \partial H[X(\mathcal{J}|\mathcal{J}')]/\partial Q^j.$$

(12)

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

The basic information theory 'regression equations' for the system of figures 1 and 2, driven by a set of external 'sensory' and other, internal, signal parameters $\mathbf{Q} = (Q^1, ..., Q^n)$ *that may be measured by the information source uncertainty of other information sources* is then precisely the set of equations (11) above.

Several features emerge directly from invoking this 'coevolutionary' approach.

The first involves Pettini's (2007) topological hypothesis: A fundamental change in the underlying topology of a system characterized by any free energy 'Morse Function' is a necessary condition for the kind of phase transition shown in figure 2. What seems clear from the neurological context is that a converse topological tuning of the threshold for the global broadcast phase transition is possible.

Second, there are several obvious possible dynamic patterns:

1. Setting equation (11) equal to zero and solving for stationary points gives attractor states since the noise terms preclude unstable equilibria.

2. This system may converge to limit cycle or pseudorandom 'strange attractor' behaviors in which the system seems to chase its tail endlessly within a limited venue – a kind of 'Red Queen' pathology.

3. What is converged to in both cases is not a simple state or limit cycle of states. Rather it is an equivalence class, or set of them, of highly dynamic information sources coupled by mutual interaction through crosstalk. Thus 'stability' in this structure represents particular patterns of ongoing dynamics rather than some identifiable static configuration.

We are deeply enmeshed in a highly recursive phenomenological stochastic differential equations (as in, e.g., Zhu et al. 2007), but in a dynamic rather than static manner. The objects of this dynamical system are equivalence classes of information sources, rather than simple 'stationary states' of a dynamical or reactive chemical system. The necessary conditions of the asymptotic limit theorems of communication theory have beaten the mathematical thicket back one layer.

Third, as Champagnat et al. (2006) note, shifts between the quasi-equilibria of a coevolutionary system can be addressed by the large deviations formalism. They find that the issue of dynamics drifting away from trajectories predicted by the canonical equation can be investigated by considering the asymptotic of the probability of 'rare events' for the sample paths of the diffusion.

By 'rare events' they mean diffusion paths drifting far away from the direct solutions of the canonical equation. The probability of such rare events is governed by a large deviation principle: when a critical parameter (designated ϵ) goes to zero, the probability that the sample path of the diffusion is close to a given rare path ϕ decreases exponentially to 0 with rate $\mathcal{I}(\phi)$, where the 'rate function' $\mathcal{I}$ can be expressed in terms of the parameters of the diffusion. This result, in their view, can be used to study long-time behavior of the diffusion process when there are multiple attractive singularities. Under proper conditions the most likely path followed by the diffusion when exiting a basin

of attraction is the one minimizing the rate function $\mathcal{I}$ over all the appropriate trajectories. The time needed to exit the basin is of the order $\exp(V/\epsilon)$ where V is a quasi-potential representing the minimum of the rate function $\mathcal{I}$ over all possible trajectories.

An essential fact of large deviations theory is that the rate function $\mathcal{I}$ which Champagnat et al. invoke can almost always be expressed as a kind of entropy, that is, having the canonical form

$$\mathcal{I} = -\sum_j P_j \log(P_j)$$

(13)

for some probability distribution. This result goes under a number of names; Sanov's Theorem, Cramer's Theorem, the Gartner-Ellis Theorem, the Shannon-McMillan Theorem, and so forth (Dembo and Zeitouni, 1998).

These considerations lead very much in the direction of equation (11), but now seen as subject to internally-driven large deviations *that are themselves described as information sources*, providing $Q = f(\mathcal{I})$-parameters that can trigger punctuated shifts between quasi-stable modes. Thus both external signals, characterized by the information source Z, and internal 'ruminations', characterized by the information source $\mathcal{I}$, can provide Q-parameters that serve to drive the system to different quasi-equilibrium 'conscious attention states' in a highly punctuated manner, if they are of sufficient magnitude.

7 Discussion and conclusions

A tuning theorem variant of the Shannon Coding Theorem that expresses the no free lunch restriction allows construction of a version of Bernard Baars' global workspace/global broadcast model of animal consciousness. Consciousness, via the giant component linking unconscious cognitive modules, and inattentional blindness, via the no free lunch condition, emerge directly, and the effects of external signals and internal ruminations can be incorporated through standard arguments leading to punctuated threshold detection.

The central evolutionary process leading to this elaborate mechanism is that the spandrel of crosstalk between unconscious cognitive modules becomes a sufficient condition for evolutionary exaptation into the arch of animal consciousness through the information theory chain rule that implies it takes more metabolic free energy to prevent correlation than to allow it. Consciousness, in terms of a rapidly shifting global neural broadcast operating in the 100 millisecond range, is not, however, necessary, as a vast spectrum of living things, both past and present, testify.

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

The parallel argument is, of course, that the similar necessary ubiquity of noise in neural process has been exapted into mechanisms of stochastic resonance amplification at various scales.

It should be obvious that roughly similar evolutionary exaptations would be available under a broad variety of astrobiological circumstances, via the statistical regularities imposed by the asymptotic limit theorems of information theory.

8 Acknowledgments

The author thanks R.G. Wallace for useful discussions.

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

9 References

Ash, R., 1990, Information Theory, Dover, New York.

Atlan, H., I. Cohen, 1998, Immune information, self organization, and meaning, International Immunology, 10:711-717.

Baars, B., 1988, A Cognitive Theory of Consciousness, Cambridge University Press, New York.

Baars, B., 2005, Global workspace theory of consciousness: toward a cognitive neuroscience of human experience, Progress in Brain Research, 150:45-53.

Baars, B., S. Franklin, 2003, How conscious experience and working memory interact, Trends in Cognitive Science, 7:166-172.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, D. Hwang, 2006, Complex networks: structure and dynamics, Physics Reports, 424:175-208.

Champagnat, N., R. Ferriere, S. Meleard, 2006, Unifying evolutionary dynamics: from individual stochastic process to macroscopic models, Theoretical Population Biology, 69:297-321.

Cover, T., J. Thomas, 2006, Elements of Information Theory, Second Edition, Wiley, New York.

Dembo, A., O. Zeitouni, 1998, Large Deviations and Applications, Springer, New York.

Dretske, F., 1994, The explanatory role of information, Philosophical Transactions of the Royal Society A, 349:59-70.

El Gamal, A., Y. Kim, 2010, Lecture Notes on Network Information Theory, ArXiv:1001.3404v4.

English, T., 1996, Evaluation of evolutionary and genetic optimizers: no free lunch. In Evolutionary Programming V: Proceedings of the Fifth Annual Conference on Evolutionary Programming, Fogel, L., P. Angeline, T. Back (eds.): 163-169, MIT Press, Cambridge, MA.

Erdos, P., A. Renyi, 1960, On the evolution of random graphs, Magyar Tud. Akad. Mat. Kutato Int. Kozl, 5:17-61.

Feynman, R., 2000, Lectures on Computation, Westview Press, New York.

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

Glazebrook, J.F., R. Wallace, 2009, Rate distortion manifolds as model spaces for cognitive information, Informatica, 33:309-346.

Gluckman, B., T. Netoff, E. Neel, W. Ditto, M. Spano, S. Schiff, 1996, Stochastic resonance in a neuronal network from mammalian brain, Physical Review Letters, 77:4098-4101.

Gould, S., R. Lewontin, 1979, The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme, Proceedings of the Royal Society of London, B, 205:581-598.

Gould, S., 2002, The Structure of Evolutionary Theory, Harvard University Press, Cambridge, MA.

Kawaguchi, M., H. Mino, D. Durand, 2011, Stochastic resonance can enhance information transmission in neural networks, IEEE Transactions on Biomedical Engineering, 58:(7) DOI 10.1109/TBME.2011.2126571.

Khinchin, A., 1957, The Mathematical Foundations of Information Theory, Dover, New York.

Park, J., P. Neelakanta, 1996, Information-theoretic aspects of neural stochastic resonance, Complex Systems, 10:55-71.

Pettini, M., 2007, Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics, Springer, New York.

Spenser, J., 2010, The giant component: the golden anniversary, Notices of the AMS, 57:720-724.

Wallace, R., 2000, Language and coherent neural amplification in hierarchical systems: renormalization and the dual information source of a generalized spatiotemporal stochastic resonance, International Journal of Bifurcation and Chaos, 10:493-502.

Wallace, R., 2005, Consciousness: A Mathematical Treatment of the Global Neuronal Workspace Model, Springer, New York.

Wallace, R., 2007, Culture and inattentional blindenss: a global workspace perspective, Journal of Theoretical Biology, 245:378-390.

Wallace, R., 2008, Toward formal models of biologically inspired, highly parallel machine cognition, International Journal of Parallel, Emergent, and Distributed Systems, 23:367-408.

Wallace, R., 2009, Programming coevolutionary machines: the emerging conundrum, International Journal of Parallel, Emergent, and Distributed Systems, 24:443-453.

Wallace, R., 2010, Tunable epigenetic catalysis: programming real-time cognitive machines, International Journal of Parallel, Emergent, and Distributed Systems, 25:209-222.

Wallace, R., 2011, Hunter-gatherers in a howling wilderness: neoliberal capitalism as a language that speaks itself,

http://precedings.nature.com/documents/5650/version/1

Wallace, R., M. Fullilove, 2008, Collective Consciousness and its Discontents, Springer, New York.

Ward, L., 2009, Physics of neural synchronation mediated by stochastic resonance, Contemporary Physics, 50:563-574.

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

Wolpert, D., W. MacReady, 1995, No free lunch theorems for search, Santa Fe Institute, SFI-TR-02-010.

Wolpert, D., W. MacReady, 1997, No free lunch theorems for optimization, IEEE Transactions on Evolutionary Computation, 1:67-82.

Zhu, R., A. Rebirio, D. Salahub, S. Kaufmann, 2007, Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models, Journal of Theoretical Biology, 246:725-745.

10 Mathematical Appendix

Messages from an information source, seen as symbols x_j from some alphabet, each having probabilities P_j associated with a random variable X, are 'encoded' into the language of a 'transmission channel', a random variable Y with symbols y_k, having probabilities P_k, possibly with error. Someone receiving the symbol y_k then retranslates it (without error) into some x_k, which may or may not be the same as the x_j that was sent.

More formally, the message sent along the channel is characterized by a random variable X having the distribution

$$P(X = x_j) = P_j, j = 1, ..., M.$$

The channel through which the message is sent is characterized by a second random variable Y having the distribution

$$P(Y = y_k) = P_k, k = 1, ..., L.$$

Let the joint probability distribution of X and Y be defined as

$$P(X = x_j, Y = y_k) = P(x_j, y_k) = P_{j,k}$$

and the conditional probability of Y given X as

$$P(Y = y_k|X = x_j) = P(y_k|x_j).$$

Then the Shannon uncertainty of X and Y independently and the joint uncertainty of X and Y together are defined respectively as

$$H(X) = -\sum_{j=1}^{M} P_j \log(P_j)$$

$$H(Y) = -\sum_{k=1}^{L} P_k \log(P_k)$$

$$H(X,Y) = -\sum_{j=1}^{M}\sum_{k=1}^{L} P_{j,k}\log(P_{j,k}).$$

(14)

The *conditional uncertainty* of Y given X is defined as

$$H(Y|X) = -\sum_{j=1}^{M}\sum_{k=1}^{L} P_{j,k}\log[P(y_k|x_j)]$$

(15)

For any two stochastic variates X and Y, $H(Y) \geq H(Y|X)$, as knowledge of X generally gives some knowledge of Y. Equality occurs only in the case of stochastic independence.

Since $P(x_j, y_k) = P(x_j)P(y_k|x_j)$, we have

$$H(X|Y) = H(X,Y) - H(Y)$$

The information transmitted by translating the variable X into the channel transmission variable Y – possibly with error – and then retranslating without error the transmitted Y back into X is defined as

$$I(X|Y) \equiv H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$$

(16)

Again, see Ash (1990), Cover and Thomas (2006) or Khinchin (1957) for details. The essential point is that if there is no uncertainty in X given the channel Y, then there is no loss of information through transmission. In general this will not be true, and herein lies the essence of the theory.

Given a fixed vocabulary for the transmitted variable X, and a fixed vocabulary and probability distribution for the channel Y, we may vary the probability distribution of X in such a way as to maximize the information sent. The capacity of the channel is defined as

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

$$C \equiv \max_{P(X)} I(X|Y)$$

(17)

subject to the subsidiary condition that $\sum P(X) = 1$.

The critical trick of the Shannon Coding Theorem for sending a message with arbitrarily small error along the channel Y at any rate $R < C$ is to encode it in longer and longer 'typical' sequences of the variable X; that is, those sequences whose distribution of symbols approximates the probability distribution $P(X)$ above which maximizes C.

If $S(n)$ is the number of such 'typical' sequences of length n, then

$$\log[S(n)] \approx nH(X)$$

where $H(X)$ is the uncertainty of the stochastic variable defined above. Some consideration shows that $S(n)$ is much less than the total number of possible messages of length n. Thus, as $n \to \infty$, only a vanishingly small fraction of all possible messages is meaningful in this sense. This observation, after some considerable development, is what allows the Coding Theorem to work so well. In sum, the prescription is to encode messages in typical sequences, which are sent at very nearly the capacity of the channel. As the encoded messages become longer and longer, their maximum possible rate of transmission without error approaches channel capacity as a limit. Again, the standard references provide details.

This approach can be, in a sense, inverted to give a 'tuning theorem' variant of the coding theorem.

Telephone lines, optical wave guides and the tenuous plasma through which a planetary probe transmits data to earth may all be viewed in traditional information-theoretic terms as a *noisy channel* around which we must structure a message so as to attain an optimal error-free transmission rate.

Telephone lines, wave guides and interplanetary plasmas are, relatively speaking, fixed on the timescale of most messages, as are most sociogeographic networks. Indeed, the capacity of a channel, is defined by varying the probability distribution of the 'message' process X so as to maximize $I(X|Y)$.

Suppose there is some message X so critical that its probability distribution must remain fixed. The trick is to fix the distribution $P(x)$ but *modify the channel* – i.e., tune it – so as to maximize $I(X|Y)$. The *dual* channel capacity C^* can be defined as

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

$$C^* \equiv \max_{P(Y),P(Y|X)} I(X|Y)$$

(18)

But

$$C^* = \max_{P(Y),P(Y|X)} I(Y|X)$$

since

$$I(X|Y) = H(X) + H(Y) - H(X,Y) = I(Y|X).$$

Thus, in a purely formal mathematical sense, *the message transmits the channel*, and there will indeed be, according to the Coding Theorem, a channel distribution $P(Y)$ which maximizes C^*.

One may do better than this, however, by modifying the channel matrix $P(Y|X)$. Since

$$P(y_j) = \sum_{i=1}^{M} P(x_i)P(y_j|x_i),$$

$P(Y)$ is entirely defined by the channel matrix $P(Y|X)$ for fixed $P(X)$ and

$$C^* = \max_{P(Y),P(Y|X)} I(Y|X) = \max_{P(Y|X)} I(Y|X).$$

Calculating C^* requires maximizing the complicated expression

$$I(X|Y) = H(X) + H(Y) - H(X,Y)$$

which contains products of terms and their logs, subject to constraints that the sums of probabilities are 1 and each probability is itself between 0 and 1. Maximization is done by varying the channel matrix terms $P(y_j|x_i)$ within the constraints. This is a difficult problem in nonlinear optimization. However, for the special case $M = L$, C^* may be found by inspection:

If $M = L$, then choose

$$P(y_j|x_i) = \delta_{j,i}$$

where $\delta_{i,j}$ is 1 if $i = j$ and 0 otherwise. For this special case

$$C^* \equiv H(X)$$

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011

with $P(y_k) = P(x_k)$ for all k. *Information is thus transmitted without error when the channel becomes 'typical' with respect to the fixed message distribution* $P(X)$.

If $M < L$ matters reduce to this case, but for $L < M$ information must be lost, leading to Rate Distortion limitations.

Thus modifying the channel may be a far more efficient means of ensuring transmission of an important message than encoding that message in a 'natural' language which maximizes the rate of transmission of information on a fixed channel.

We have examined the two limits in which either the distributions of $P(Y)$ or of $P(X)$ are kept fixed. The first provides the usual Shannon Coding Theorem, and the second a tuning theorem variant, i.e. a tunable, retina-like, Rate Distortion Manifold, in the sense of Glazebrook and Wallace (2009).

Nature Precedings : hdl:10101/npre.2011.6115.1 : Posted 14 Jul 2011