Version 3.1

Consciousness: A Simple Information Theory Global Workspace Model

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

April 15, 2011

Abstract

The asymptotic limit theorems of information theory permit a concise formulation of Bernard Baars' global workspace/global broadcast picture of consciousness, focusing on how networks of unconscious cognitive modules are driven by the classic 'no free lunch' argument into shifting, tunable, alliances having variable thresholds for signal detection. The model directly accounts for the punctuated characteristics of many conscious phenomena, and derives the inherent necessity of inattentional blindness and related effects.

Key Words: asymptotic limit theorem, ergodic, network topology, no free lunch, phase transition

1 Introduction: Cognition as 'language'

A perhaps oversuccinct summary of Baars' global workspace model of 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 (e.g., Baars, 1988, 2005; Baars and Franklin, 2003). We can uncover much of this basic mechanism from a remarkably simple application of the asymptotic limit theorems of information theory, once a broad range of cognitive processes is recognized as inherently characterized by ergodic information sources – generalized languages, if you will (Wallace, 2000). This allows mapping physiological unconscious cognitive modules onto an abstract network of interacting information sources, permitting a simplified mathematical attack based on phase transitions in network topology.

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

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 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

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

(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 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

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

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

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

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.

2 No free lunch: a little information theory

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

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

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

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

(2)

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

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

(3)

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

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

(4)

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

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

(5)

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

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

(6)

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

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

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). These results can be used to directly derive the famous 'no free lunch' theorem of Wolpert and Macready (1995, 1997). As English (1996) states the matter,

> ...Wolpert and Macready... have established that there exists no generally superior 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 argument of this section, there will be a channel coding of Y which, when properly tuned, is most efficiently transmitted 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 problems, although there will likely be an optimal

algorithm for any given set.

3 Dynamic networks of unconscious cognitive modules

Based on the no free lunch argument of the previous section, it is clear that different challenges facing a conscious entity must be met be different arrangements of basic cognitive faculties. It is now possible to make a very abstract picture of the brain, not based on its anatomy, but rather on the linkages between the information sources dual to the basic physiological and learned unconscious cognitive modules (UCM) that form Baars' global workspace/global broadcast. That is, *the remapped brain network is reexpressed in terms of the information sources dual to the UCM.* Given two distinct problems classes (e.g., playing tennis vs. interacting with a significant other), 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), as done in Wallace (2011).

The emergence of a closely linked set of information sources dual to the UCM into a global workspace/broadcast system itself depends 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. If $c < 1$ all the linked

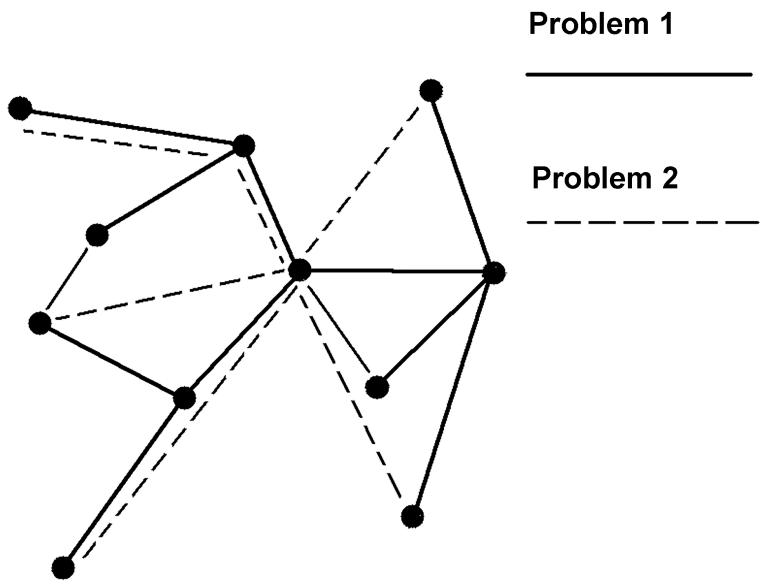

Figure 1: By the no free lunch theorem, two markedly different problems will be optimally solved by two different linkages of available unconscious cognitive modules into different temporary global workspace/broadcast networks, here represented by crosstalk among their dual information sources rather than the physiological UCM themselves.

subnetworks are very small, *and no global broadcast can take place*. If $c = 1$ there is a single large interlinked component of a size $\approx n^{2/3}$. 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

$$\exp(-cy) = 1 - y. \tag{7}$$

Then

$$y = \frac{W(-c/\exp(c)) + c}{c}, \tag{8}$$

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 – the global broadcast for interacting UCM. To the left of $c = 1$ there is no giant component, and large scale – i.e., conscious – 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 – the emergence of consciousness – is dependent on a network topology that may itself be tunable.

According to this argument, if the network topology becomes tuned, then a sensory input parameterized by c with $c < 1$ can trigger a global broadcast.

One imagines a set of sensory inputs, $C = \{c_1, ..., c_j\}$ affecting a highly multidimensional structure of interacting UCM, represented abstractly here by the network of their dual information sources. If the set is tuned by the no free lunch theorem to maximize response to the 'problem' defined by a particular c_i, so that $c_i \ll 1$ can trigger a global broadcast, then the other sensory inputs will be inherently subject to inattentional blindness, a somewhat simpler picture than that presented by Wallace (2007).

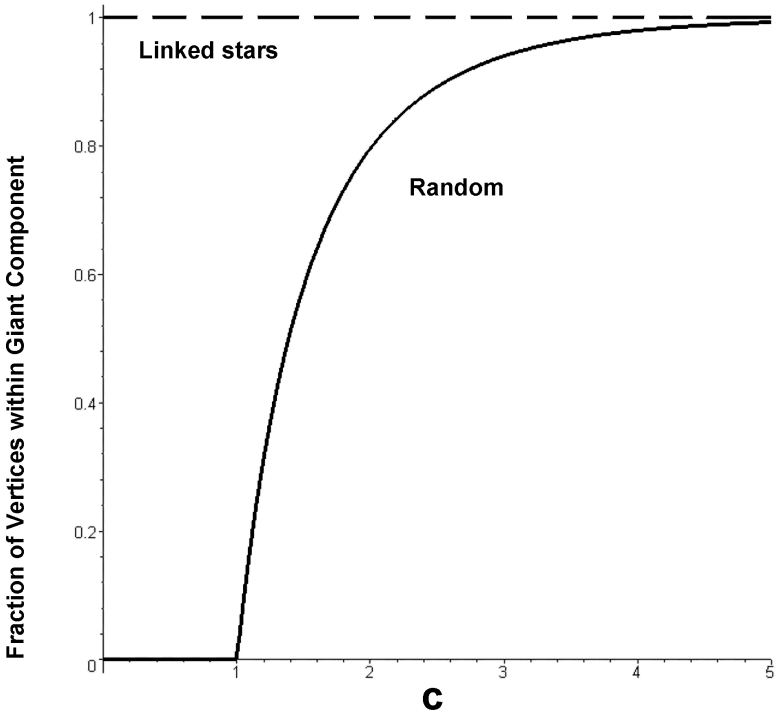

Figure 2: Fraction of network nodes in the giant component as a function of the coupling parameter c. The solid line represents a random graph, the dotted line a star-of-stars-of-stars network in which all nodes are interconnected, showing that the dynamics of giant component emergence are highly dependent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

4 Discussion and conclusions

An elementary tuning theorem variant of the Shannon Coding Theorem that expresses the no free lunch argument allows construction of a simple version of Bernard Baars' global workspace/global broadcast model of consciousness. Punctuated accession to consciousness, via sudden onset of a giant component, and inattentional blindness, via the no free lunch restriction, emerge directly. More complicated models are required to explore the nature of the phase transition implied by the solid line in figure 2 (Wallace, 2005), the effects of embedding culture on inattentional blindness (Wallace, 2007), and the conundrum presented by institutional or machine versions of consciousness that can support multiple, interacting, global broadcasts (Wallace and Fullilove, 2008; Wallace, 2008, 2009, 2010).

5 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.

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

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.

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

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

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.