Version 18

Lurching Toward Chernobyl

Dysfunctions of Real-Time Computation

Rodrick Wallace, Ph.D.
Division of Epidemiology
The New York State Psychiatric Institute*

November 20, 2008

Abstract

Cognitive biological structures, social organizations, and computing machines operating in real time are subject to Rate Distortion Theorem constraints driven by the homology between information source uncertainty and free energy density. This exposes the unitary structure/environment system to a relentless entropic torrent compounded by sudden large deviations causing increased average distortion between intent and impact, particularly as demands escalate. The phase transitions characteristic of information phenomena suggest that, rather than graceful decay under increasing load, these structures will undergo punctuated degradation akin to spontaneous symmetry breaking in physical systems. Rate distortion problems, that also affect internal structural dynamics, can become synergistic with limitations equivalent to the inattentional blindness of natural cognitive processes. These mechanisms, and their interactions, are unlikely to scale well, so that, depending on architecture, enlarging the structure or its duties may lead to a crossover point at which added resources must be almost entirely devoted to ensuring system stability – a form of allometric scaling familiar from biological examples. This suggests a critical need to tune architecture to problem type and system demand. A real-time computational structure and its environment are a unitary phenomenon, and environments are usually idiosyncratic. Thus the resulting path dependence in the development of pathology could often require an individualized approach to remediation more akin to an arduous psychiatric intervention than to the traditional engineering or medical quick fix. Failure to recognize the depth of these problems seems likely to produce a relentless chain of the Chernobyl-like failures that are necessary, but often insufficient, for remediation under our system.

Key Words cognition, entropy, free energy, groupoid, information theory, large deviations, Morse Theory, no free lunch, parallel, pathology, rate distortion, scaling laws, spontaneous symmetry breaking

*Address correspondence to Rodrick Wallace, 549 W. 123 St., New York, NY, 10027 USA, email wallace@pi.cpmc.columbia.edu.

1 Introduction

Massively parallel machines are currently used or proposed for the regulation of dynamic processes in real time. Potential critical applications include financial systems and communications networks, refineries, nuclear reactors, chemical factories, large scale traffic control, the piloting of individual vehicles, and so on.

Many physiological and psychological systems in higher animals, including man, are both cognitive in the Atlan/Cohen sense (Atlan and Cohen, 1998), and also operate rapidly enough to be classified as real-time. These include immune and blood pressure regulation mechanisms, cognitive gene expression, and higher order cognition and consciousness (e.g., Baars, 2005; Dretske, 1981, 1988, 1994; Wallace, 2005; Wallace and Wallace, 2008).

Social structures ranging from animal hive colonies to modern giant corporations, their institutional networks of competition and cooperation, and the associated political empires which they dominate, engage in elaborate processes of distributed cognition that must also operate in real-time (e.g., Wallace and Fullilove, 2008).

Here we will examine limits placed on real-time systems by the Rate Distortion Theorem and by homologies with thermodynamic free energy, and the manner in which those limits can produce characteristic patterns of system degradation and failure.

We begin with a highly formal description of phase transitions in cognitive systems, extending perspectives from physical theory to 'necessary conditions' statistical models of cognitive process based on the asymptotic limits of information theory.

Landau's famous insight regarding phase change in physical systems was that second order phase transitions are usually in the context of a significant alteration in symmetry, with one phase being far more symmetric than the other (e.g., Pettini, 2007; Landau and Lifshitz, 2007). A symmetry is lost in the transition, a phenomenon called spontaneous symmetry breaking. The greatest possible set of symmetries in a physical system is that of the Hamiltonian describing its energy states. Usually states accessible at lower temperatures will lack the symmetries available at higher temperatures, so that the lower temperature phase is less symmetric: The randomization of higher temperatures ensures that higher symmetry/energy states will then be accessible to the system. A change in symmetry must, of necessity, be discontinuous, so that lowering temperatures inevitably leads to punctuated transitions in such systems. One can indeed construct a good phenomenological model using group representations (Pettini, 2007).

What of biological and cognitive structures that cannot easily be described using elementary physical models or simple group symmetries? What of systems where the physical temperature is not the determining factor in punctuated change?

We will be concerned here with systems having associated information sources that can be described in terms of groupoids, a natural generalization of groups described in the Appendix that is finding increasingly widespread use in biology and cognitive theory (e.g., Golubitsky and Stewart, 2006). In particular, as we argue below, a broad swath of cognitive phenomena can be characterized in terms of information sources (e.g., Wallace, 2005; Fullilove and Wallace, 2008).

Mathematical preliminaries require a brief exploration of the homology between information and free energy.

2 Free Energy Density and Information Source Uncertainty

Information source uncertainty can be defined in several ways. Khinchin (1957) describes the fundamental 'E-property' of a stationary, ergodic information source as the ability, in the limit of infinity long output, to classify strings into two sets;

[1] a very large collection of gibberish which does not conform to underlying rules of grammar and syntax, in a large sense, and which has near-zero probability, and

[2] a relatively small 'meaningful' set, in conformity with underlying structural rules, having very high probability.

The essential content of the Shannon-McMillan Theorem is that, if $N(n)$ is the number of 'meaningful' strings of length n, then the uncertainty of an information source X can be defined as

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

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

$$\lim_{n \to \infty} \frac{H(X_0, ..., X_n)}{n+1},$$

(1)

where $H(...|...)$ and $H(....)$ are conditional and joint Shannon uncertainties defined from the appropriate cross-sectional string probabilities (e.g., Ash, 1990; Cover and Thomas, 1991; Khinchin, 1957).

Information source uncertainty has an important heuristic interpretation. Ash (1990) puts it this way:

> ...[W]e may regard a portion of text in a particular language as being produced by an information source. The probabilities $P[X_n = a_n | X_0 = a_0, ...X_{n-1} = a_{n-1}]$ may be estimated from the available data about the language; in this way we can estimate the uncertainty associated with the language. A large uncertainty means, by the [Shannon-McMillan Theorem], a large number of 'meaningful' sequences. Thus given two languages with uncertainties H_1 and H_2 respectively, if $H_1 > H_2$, then in the absence of noise it is easier to communicate in the first language; more can be said in the same amount of time. On the other hand, it will be easier to reconstruct a scrambled portion of text in the second language, since fewer of the possible sequences of length n are meaningful.

The free energy density of a physical system having volume V and partition function $Z(K)$ derived from the system's Hamiltonian at inverse temperature K is (e.g., Landau and Lifshitz 2007)

$$F[K] = \lim_{V \to \infty} -\frac{1}{K} \frac{\log[Z(K, V)]}{V} =$$

$$\lim_{V \to \infty} \frac{\log[\hat{Z}(K, V)]}{V},$$

(2)

where $\hat{Z} = Z^{-1/K}$.

Feynman (2000), following Bennett, concludes that the information contained in a message is simply the free energy needed to erase it. The argument is direct. Thus, and accordingly, information source uncertainty is homologous to free energy density as defined above.

3 Groupoid Free Energy

Equivalence classes define groupoids, by the mechanisms described in the Mathematical Appendix. The basic equivalence classes of a cognitive or biological structure will define the basic transitive groupoids, and higher order systems can be constructed by the union of these transitive groupoids, having larger alphabets that allow more complicated statements in the sense of Ash above. We associate information sources with transitive groupoids, and with the larger groupoids constructed from them. The more complicated the groupoid, the greater the information source uncertainty, following Ash's reasoning.

Given an appropriately scaled, dimensionless, inverse embedding temperature-analog K, we propose that the probability of an information source H_{G_i}, representing some groupoid element G_i, will be given by the classic relation (e.g., Landau and Lifshitz 2007)

$$P[H_{G_i}] = \frac{\exp[-H_{G_i}K]}{\sum_j \exp[-H_{G_j}K]}, \quad (3)$$

where the normalizing sum is appropriately over all possible elements of the largest available symmetry groupoid. By the arguments above, compound sources, formed by the union of (interaction of elements from) underlying transitive groupoids, being more complex, will all have higher free-energy-density-equivalents than those of the base (transitive) groupoids.

Let

$$Z_G \equiv \sum_j \exp[-H_{G_j}K]. \quad (4)$$

We now define the *Groupoid Free Energy* (GFE) of the system, F_G, at inverse normalized equivalent temperature K, as

$$F_G[K] \equiv -\frac{1}{K}\log[Z_G[K]]. \quad (5)$$

We have expressed the probability of an information source in terms of its relation to a fixed, appropriately normalized, inverse system temperature. This gives a statistical thermodynamic means of defining a 'higher' free energy construct – $F_G[K]$ – to which we can now apply Landau's fundamental heuristic phase transition argument (Landau and Lifshitz 2007; Skierski et al. 1989; Pettini 2007).

Absent a high value of the temperature-equivalent, in this model, only the simplest transitive groupoid structures can be manifest. A full treatment from this perspective requires invocation of groupoid representations, no small matter (e.g., Bos, 2007; Buneci, 2003).

Somewhat more rigorously, the elaborate renormalization schemes of Wallace (2005) may now be imposed on $F_G[K]$ itself, leading to a spectrum of highly punctuated transitions in the overall system of information sources. The essential point is that $F_G[K]$ is unlikely to scale with a renormalization transform as simply as does physical free energy, and this leads to very complicated 'biological' renormalization strategies. See Wallace (2005), Wallace and Fullilove, (2008) or Wallace et al., (2007) for details.

Most deeply, however, an extended version of Pettini's (2007) Morse-Theory-based topological hypothesis can now be invoked, i.e., that changes in underlying groupoid structure are a necessary (but not sufficient) consequence of phase changes in $F_G[K]$. Necessity, but not sufficiency, is important, as it allows for mixed symmetries. An outline of the theory is presented in the Appendix. For details see, e.g., Matsumoto (2002) or Pettini (2007).

As the temperature-analog declines, in this model, the system can undergo punctuated groupoid symmetry reductions representing fundamental phase transitions.

Dynamical behavior away from critical points will be determined, in this model, by Generalized Onsager Relations, also explored more fully in the Appendix.

We next apply this formalism to examples of of purely internal, and of linked internal-external, cognitive function.

4 High Order Cognition

According to Atlan and Cohen (1998), the essence of cognition is comparison of a perceived external signal with an internal, learned picture of the world, and then, upon that comparison, the choice of one response from a much larger repertoire of possible responses. Such reduction in uncertainty inherently carries information, and, following Wallace (2000, 2005) it is possible to make a very general model of this process as an information source.

A pattern of 'sensory' input, say an ordered sequence $y_0, y_1, ...,$ is mixed in a systematic (but unspecified) algorithmic manner with internal 'ongoing' activity, the sequence $w_0, w_1, ...,$ to create a path of composite signals $x = a_0, a_1, ..., a_n, ...,$ where $a_j = f(y_j, w_j)$ for some function f. This path is then fed into a highly nonlinear, but otherwise similarly unspecified, decision oscillator which generates an output $h(x)$ that is an element of one of two disjoint sets B_0 and B_1. We take

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

$$B_1 \equiv b_{k+1}, ..., b_m. \quad (6)$$

Thus we permit a graded response, supposing that if

$$h(x) \in B_0$$

(7)

the pattern is not recognized, and if

$$h(x) \in B_1$$

(8)

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

Our focus is on those composite paths x that trigger pattern recognition-and-response. That is, given a fixed initial state a_0, such that $h(a_0) \in B_0$, 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 $0 \leq j < m$, but $h(a_0, ..., a_m) \in B_1$.

For each positive integer n, let $N(n)$ be the number of grammatical and syntactic high probability paths of length n which begin with some particular a_0 having $h(a_0) \in B_0$ and lead to the condition $h(x) \in B_1$. We shall call such paths meaningful and assume $N(n)$ to be considerably less than the number of all possible paths of length n – pattern recognition-and-response is comparatively rare. We again assume that the longitudinal finite limit $H \equiv \lim_{n \to \infty} \log[N(n)]/n$ both exists and is independent of the path x. We will – not surprisingly – call such a cognitive process *ergodic.*

Note that disjoint partition of state space may be possible according to sets of states which can be connected by meaningful paths from a particular base point, leading to a natural coset algebra of the system, a groupoid. This is a matter of some importance.

It is thus possible to define an 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 may be defined which satisfy the relations of equation (1) above.

This information source is taken as *dual* to the ergodic cognitive process.

Dividing the full set of possible responses into the sets B_0 and B_1 may itself require higher order cognitive decisions by another module or modules, suggesting the necessity of choice within a more or less broad set of possible quasi-languages. This would directly reflect the need to shift gears according to the different challenges faced by the organism, machine, or social group.

'Meaningful' paths – creating an inherent grammar and syntax – have been defined entirely in terms of system response, as Atlan and Cohen (1998) propose. This formalism can easily be applied to the stochastic neuron in a neural network, as done in Wallace (2005).

A formal equivalence class algebra can be constructed for a cognitive process characterized by a dual information source by choosing different origin points a_0, in the sense above, and defining equivalence of two states by the existence of a high-probability meaningful path connecting them with the same origin. Disjoint partition by equivalence class, analogous to orbit equivalence classes for dynamical systems, defines the vertices of a network of cognitive dual languages. Each vertex then represents a different information source dual to a cognitive process. This is not a direct representation as in a neural network, or of some circuit in silicon. It is, rather, an abstract set of 'languages' dual to the cognitive processes instantiated by biological structures, machines, social process, or their hybrids. Our particular interest, however, is in an interacting network of cognitive processes.

This structure generates a groupoid, in the sense of the Appendix. Recall that states a_j, a_k in a set A are related by the groupoid morphism if and only if there exists a high-probability grammatical path connecting them to the same base point, and tuning across the various possible ways in which that can happen – the different cognitive languages – parametizes the set of equivalence relations and creates the groupoid.

We now envision an average mean field mutual information linking different information sources associated with the transitive groupoids defined by this network. Call that mean field $\mathcal{I}$. Another possible interpretation is of an average probability of nondisjuctive 'weak' ties $\mathcal{P}$ (*sensu* Granovetter, 1973) linking the different ergodic dual information sources. Then, for the Groupoid Free Energy calculation above, take $K \propto 1/\mathcal{I}, 1/\mathcal{P}$. Increasing $\mathcal{I}$ or $\mathcal{P}$ then, increases the linkage across the transitive groupoids of the cognitive system, leading, in a highly punctuated way, to larger and larger processes of collective cognition using progressively larger 'alphabets' and having, in the sense of Ash above, progressively larger values of the associated dual information source.

A second model arises in a natural manner by taking $1/K$ as the *mean number*, $\mathcal{N}$, of linkages between dual information sources in the abstract network. This leads to generalizations of the Erdos/Renyi random network formalism, and its inherent phase transitions.

Both approaches can be extended to second order as an analog to hierarchical regression. The first generalization is via a kind of universality class tuning, and the second by means of a renormalization in which couplings at or above a tunable limit are set to 1 and those below to 0. A Morse Theory topological tuning results directly from the latter approach. Evolutionary process, or engineering design, are not necessarily restricted, however, to these two exactly solvable models.

Wallace (2005) and Wallace and Fullilove (2008) use simplified forms of this argument to characterize consciousness and distributed institutional cognition, respectively. Our particular interest, however, is in the ways such cognitive structures respond to challenges in real time: individual and institutional distributed cognition do not occur in a vacuum, but in the context of demands for prompt action from an embedding ecological structure. We will attempt to characterize pathologies of such real time response. To do this we must iterate the argument.

5 Real Time Systems

5.1 Cognitive Dynamics in Real Time

Real time problems are inherently rate distortion problems: The implementation of a complex cognitive structure, say a sequence of control orders generated by some dual information source Y, having output $y^n = y_1, y_2, ...$ is 'digitized' in terms of the observed behavior of the regulated system, say the sequence $b^n = b_1, b_2, ...$ The b_i are thus what happens in real time, the actual impact of the cognitive structure on its embedding environment. Assume each b^n is then deterministically retranslated back into a reproduction of the original control signal,

$$b^n \to \hat{y}^n = \hat{y}_1, \hat{y}_2, ...$$

Define a distortion measure $d(y, \hat{y})$ which compares the original to the retranslated path. Suppose that with each path y^n and b^n-path retranslation into the y-language, denoted $\hat{y}^n$, there are associated individual, joint, and conditional probability distributions

$$p(y^n, p(\hat{y}^n), p(y^n|\hat{y}^n).$$

The average distortion is defined as

$$D \equiv \sum_{y^n} p(y^n) d(y^n, \hat{y}^n).$$

(9)

It is possible, using the distributions given above, to define the information transmitted from the incoming Y to the outgoing $\hat{Y}$ process using the Shannon source uncertainty of the strings:

$$I(Y, \hat{Y}) \equiv H(Y) - H(Y|\hat{Y}) = H(Y) + H(\hat{Y}) - H(Y, \hat{Y}).$$

If there is no uncertainty in Y given the retranslation $\hat{Y}$, then no information is lost, and the regulated system is perfectly under control.

In general, this will not be true.

The *information rate distortion function* $R(D)$ for a source Y with a distortion measure $d(y, \hat{y})$ is defined as

$$R(D) = \min_{p(y,\hat{y}); \sum_{(y,\hat{y})} p(y)p(y|\hat{y})d(y,\hat{y}) \leq D} I(Y, \hat{Y}).$$

(10)

The minimization is over all conditional distributions $p(y|\hat{y})$ for which the joint distribution $p(y, \hat{y}) = p(y)p(y|\hat{y})$ satisfies the average distortion constraint (i.e., average distortion $\leq D$).

The *Rate Distortion Theorem* states that $R(D)$ *is the minimum necessary rate of information transmission which ensures the transmission does not exceed average distortion* D. Thus $R(D)$ defines a minimum necessary channel capacity. Cover and Thomas (1991) or Dembo and Zeitouni (1998) provide details. The rate distortion function has been calculated for a number of systems. Cover and Thomas (1991, Lemma 13.4.1) show that $R(D)$ is necessarily a decreasing convex function of D, that is, always a reverse J-shaped curve. For a Gaussian channel having noise with zero mean and variance σ^2,

$$R(D) = 1/2 \log[\sigma^2/D], 0 \leq D \leq \sigma^2$$

$$R(D) = 0, D > \sigma^2.$$

(11)

Recall, now, the relation between information source uncertainty and channel capacity (e.g., Ash, 1990; Cover and Thomas, 1991):

$$H[\mathbf{X}] \leq C,$$

(12)

where H is the uncertainty of the source X and C the channel capacity, defined according to the relation,

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

(13)

where $P(X)$ is the probability distribution of the message chosen so as to maximize the rate of information transmission along a channel Y.

The rate distortion function $R(D)$ defines the minimum channel capacity necessary for the system to have average distortion less than or equal D, placing a limits on information source uncertainty. Thus, we suggest distortion measures can drive information system dynamics. That is, the rate distortion function also has a homological relation to free energy density.

We can model the disjunction between intent and impact of a cognitive system interacting with an embedding environment using a simple extension of the language-of-cognition approach above. Recall that cognitive processes can be formally associated with information sources, and how a formal equivalence class algebra can be constructed for a complicated cognitive system by choosing different origin points in a particular abstract 'space' and defining the equivalence of two states by the existence of a high probability meaningful path connecting each of them to some defined origin point within that space. Disjoint partition by equivalence class is analogous to orbit equivalence relations for dynamical systems, and defines the vertices of a network of cognitive dual languages available to the system: Each vertex represents a different information source dual to a cognitive process. The structure creates a large groupoid, with each orbit corresponding to a transitive groupoid whose disjoint union is the full groupoid. We can apply the spontaneous symmetry breaking argument to increasing disjunction between cognitive intent and system impact as follows:

With each element of the (large) cognitive groupoid G_i we can associate a dual information source H_{G_i}. Let $R(D)$ be the rate distortion function between the message sent by the cognitive process and the observed impact. Remember that both H_{G_i} and $R(D)$ are free energy density measures.

The essential argument is that $R(D)$ is an embedding context for the underlying cognitive process. The argument-by-abduction from physical theory is, then, that $R(D)$ constitutes a kind of thermal bath for the processes of cognition. Thus we can write the probability of the dual cognitive information source H_{G_i} as

$$P[H_{G_i}] = \frac{\exp[-H_{G_i}/\kappa R(D)]}{\sum_j \exp[-H_{G_j}/\kappa R(D)]},$$

(14)

where κ is an appropriate dimensionless constant characteristic of the particular system. The sum is over all possible elements of the largest available symmetry groupoid. By the usual arguments, compound sources, formed by the union of underlying transitive groupoids, being more complex, will have higher free-energy-density equivalents than those of the base transitive groupoids.

We can apply the Groupoid Free Energy phase transition arguments from above, remembering that the Rate Distortion Function $R(D)$ is always a decreasing convex function of D (Cover and Thomas, 1991). For real time cognitive systems, increasing average distortion between cognitive intent and observed impact will 'lower the temperature' so as to drive the cognitive process, in a highly punctuated manner, relentlessly toward simpler and less rich behaviors.

5.2 Rate Distortion Dynamics

Here we iterate the model yet again, examining the time behavior of the Rate Distortion Function itself. $R(D)$ may not simply play the passive role of a temperature-analog in real systems, but can have its own internal dynamics. These can, by the mechanisms above, drive the dynamics of the underlying cognitive structure.

Recall equations (12) and (13) and the definition of the rate distortion function from equation (10).

$R(D)$ defines the minimum channel capacity necessary for the system to have average distortion less than or equal D, placing a limits on information source uncertainty. Thus, we suggest that, *since* $R(D)$ *is itself homologous to free energy density*, distortion measures can drive rate distortion dynamics.

That is, we are led to propose, as a heuristic for real systems, that the Rate Distortion Function itself will have a dynamics that can be described using a distortion parameter. In general, take R as parametized, not only by the distortion D, but by some vector of variates $\mathbf{D} = (D_1, ..., D_k)$, for which the first component is the average distortion. The assumed dynamics are, following the general theory of section 11.2 below, then driven by gradients in the *rate distortion disorder* defined as

$$S_R \equiv R(\mathbf{D}) - \sum_{i=1}^{k} D_i \partial R/\partial D_i,$$

(15)

leading to the deterministic and stochastic systems of equations analogous to the Onsager relations of nonequilibrium thermodynamics:

$$dD_j/dt = \sum_i L_{j,i} \partial S_R/\partial D_i$$

(16)

and

$$dD_t^j = L^j(D_1, ..., D_k, t)dt + \sum_i \sigma^{j,i}(D_1, ..., D_k, t)dB_t^i,$$

(17)

where the dB_t^i represent added, often highly structured, stochastic 'noise' whose properties are characterized by their quadratic variation (e.g., Protter, 1995).

A simple Gaussian channel with noise having zero mean and variance $\sigma^2 = 1$, has a Rate Distortion function

$$R(D) = 1/2 \log[1/D],$$

so that,

$$S_R(D) = R(D) - DdR(D)/dD = 1/2 \log(1/D) + 1/2.$$

(18)

The simplest possible Onsager relation becomes

$$dD/dt \propto -dS_R/dD = \frac{1}{2D},$$

(19)

where $-dS_R/dD$ represents the force of an entropic torrent, a kind of internal dissipation inevitably driving the system toward greater distortion.

This has the solution

$$D \propto \sqrt{t}.$$

(20)

Similar results will apply to any of the reverse-J-shaped relations which must inevitably characterize $R(D)$. That is, the rate distortion function is necessarily a convex decreasing function of the average distortion D, whatever distortion measure is chosen. The implication is that real time cognitive systems, that must inevitably interact with an embedding context, will be subject to a relentless entropic force, requiring a constant energy expenditure for maintenance of some fixed average distortion D in the relation between system effort and system impact.

Absent such a contravening constraint,

$$D = f(t),$$

(21)

with $f(t)$ monotonic increasing in t.

This relation has considerable implication for the stability of the internal cognitive processes driving the real-time interaction. The mean number and field models of cognition in real time systems are, then, characterized here by a series of critical distortions D_{C_i} representing transitions to successively simpler symmetry groupoids that impose a progressive degradation on cognitive function. Absent the constant input of free energy to a real time cognitive system, then, progressive punctuated degradation in overall cognitive function is inevitable. This should apply to biological, social, and mechanical systems and their hybrids.

Application of a standard Onsager-Machlup large deviations argument suggests that D can undergo predictable large-scale excursions that may greatly challenge the stability of real time cognitive systems.

6 Large Deviations

Section 5 described how a homology with free energy allows construction of a dynamical theory for the rate distortion function associated with the mismatch between structure implementation and structure impact, using a formalism similar to the Onsager relations of nonequilibrium thermodynamics. Below we will relate the average distortion defined by that mismatch to pathological resilience shifts affecting internal structure function, as defined by the information source dual to internal cognition. Inversely, the rate distortion function, from its homology with free energy, can drive sudden, jet-like large deviations of average distortion which are finely patterned by a Hamiltonian-like function. This permits a formal classification of seemingly 'idiosyncratic' fluctuations in average distortion between structure implementation and structure impact which can, in turn, trigger internal system resilience transitions in the sense we will describe later.

Thus a broad class of seemingly-random failures of cognitive, real-time structures may, in fact, be subject to formal description and even prediction.

We first begin with a condensed review of the standard theory of large deviations.

The macroscopic behavior of a complicated physical system in time is assumed to be described by the phenomenological Onsager relations giving large-scale fluxes as

$$\sum_i W_{i,j} dK_j/dt = \partial S/\partial K_i,$$

(22)

where the $W_{i,j}$ are appropriate constants, S is the system entropy and the K_i are the generalized coordinates which parametize the system's free energy.

Entropy is defined from free energy F by a Legendre transform:

$$S \equiv F - \sum_j K_j \partial F/\partial K_j,$$

where the K_j are appropriate system parameters.

Neglecting volume problems for the moment, free energy can be defined from the system's partition function Z as

$$F(K) = -1/K \log[Z(K)].$$

The partition function Z, in turn, is defined from the system Hamiltonian – defining the energy states – as

$$Z(K) = \sum_j \exp[-KE_j],$$

where K is an inverse temperature or other parameter and the E_j are the energy states. See any good statistical mechanics text for details, e.g. Landau and Lifshitz, (2007).

Inverting the Onsager relations gives

$$dK_i/dt = \sum_j L_{i,j} \partial S/\partial K_j = L_i(K_1, ..., K_m, t) \equiv L_i(K, t).$$

(23)

The terms $\partial S/\partial K_i$ are macroscopic driving forces dependent on the entropy gradient.

Let a white Brownian noise $\epsilon(t)$ perturb the system, so that

$$dK_i/dt = \sum_j L_{i,j} \partial S/\partial K_j + \epsilon(t) = L_i(K, t) + \epsilon(t),$$

(24)

where the time averages of ϵ are $< \epsilon(t) >= 0$ and $< \epsilon(t)\epsilon(0) >= \mu\delta(t)$. $\delta(t)$ is the Dirac delta function, and we take K as a vector in the K_i.

Following Luchinsky (1997), if the probability that the system starts at some initial macroscopic parameter state K_0 at time $t = 0$ and gets to the state $K(t)$ at time t is $P(K, t)$, then a somewhat subtle development (e.g., Feller 1971) gives the forward Fokker-Planck equation for P:

$$\partial P(K, t)/\partial t = -\nabla \cdot (L(K, t)P(K, t)) + (\mu/2)\nabla^2 P(K, t).$$

(25)

In the limit of weak noise intensity this can be solved using the WKB, i.e., the eikonal, approximation, as follows. Take

$$P(K, t) = z(K, t) \exp(-s(K, t)/\mu).$$

(26)

$z(K, t)$ is a prefactor and $s(K, t)$ is a classical action satisfying the Hamilton-Jacobi equation, which can be solved by integrating the Hamiltonian equations of motion. The equation reexpresses $P(K, t)$ in the usual parametized negative exponential format.

Let $p \equiv \nabla s$. Substituting and collecting terms of similar order in μ gives

$$dK/dt = p + L,$$

$$dp/dt = -\partial L/\partial K p,$$

$$-\partial s/\partial t \equiv h(K, p, t) = pL(K, t) + \frac{p^2}{2},$$

(27)

with $h(K, t)$ the Hamiltonian for appropriate boundary conditions.

Again following Luchinsky (1997), these Hamiltonian equations have two different types of solution, depending on p. For $p = 0, dK/dt = L(K, t)$ which describes the system in the absence of noise. We expect that with finite noise intensity the system will give rise to a distribution about this deterministic path. Solutions for which $p \neq 0$ correspond to optimal paths along which the system will move with overwhelming probability.

In sum, to again paraphrase Luchinsky (1997), large fluctuations, although infrequent, are fundamental in a broad range of processes, and it was recognized by Onsager and Machlup (1953) that insight into the problem could be gained from studying the distribution of fluctuational paths along which the system moves to a given state. This distribution is a fundamental characteristic of the fluctuational dynamics, and its understanding leads toward control of fluctuations. Fluctuational motion from the vicinity of a stable state may occur along different paths. For large fluctuations, the distribution of these paths peaks sharply along an optimal, most probable, path. In the theory of large fluctuations, the pattern of optimal paths plays a role similar to that of the phase portrait in nonlinear dynamics.

The essential insight is that, by invoking the results of the section on rate distortion dynamics, we can substitute the rate distortion function for free energy in this argument and paramatize it by a number of variates including average distortion between the implementation of a cognitive system and its actual impact on the embedding environment.

From the perspective of this section, then, solutions with $p \neq 0$ correspond to optimal paths that will drive fluctuations in the average distortion between structure implementation and structure impact. As contexts, these can, we will argue, cause sudden shifts in a structure's internal resilience domain, in the sense of the development below. Thus understanding the large deviations possible to the implementation/impact mismatch would be a useful tool in predicting, mitigating, or remediating a spectrum of 'idiosyncratic' behavioral pathologies of a real time cognitive structure.

7 Extending the Model

We have proposed that an entropic torrent or sudden distortion jet can degrade both the communication between internal structural workspaces and the effect of the structure's output on the embedding context with which it interacts. We now propose that the relation between such effects will be synergistic and nonlinear, highly dependent on both architecture and demand.

If we focus entirely on the internal cognitive process or processes, then a number of 'natural' models emerge. For a single workspace internal cognitive process, the simplest generalization of equation (3) is just

$$P[H_{G_i}] = \frac{\exp[-H_{G_i}/T]}{\sum_j \exp[-H_{G_j}/T]}$$

with

$$T \propto \mathcal{P}^{\alpha} R(D)^{\beta}$$

or

$$T \propto \mathcal{N}^{\alpha} R(D)^{\beta}$$

where $\mathcal{P}$ is the mean strength of weak ties linking nodes of the internal cognitive groupoid network, $\mathcal{N}$ the mean number, and $R(D)$ the Rate Distortion function for the linkage of the internal cognitive process to its impact on the external, embedding environment.

Wallace and Fullilove (2008) describe institutional distributed cognition – effectively a collective consciousness – in terms of multiple internal cognitive workspaces, which must not only interact with an external environment, but must collaborate with other, internal, workspaces. Then we have, most simply, for this three-fold structure,

$$T \propto \mathcal{N}^{\alpha} \mathcal{R}^{\beta} R(D)^{\delta},$$

where $\mathcal{R}$ is *the average rate distortion function for communication between internal workspaces*, essentially a measure of internal bandwidth, and the exponents are all positive real numbers. Thus the dynamics are driven by interactions within and between internal global broadcasts and by the interaction between the cognitive system and its environment.

The synergistic product arises from the fact that failure at any one stage represents failure of the system, a linear chain model for which the strength of all is determined by the weakest link.

More generally, we can characterize T as a 'synergism function', monotonic-increasing in its components, and zero if any component is zero.

Declines in T drive punctuated declines in internal cognitive richness via a groupoid version of spontaneous symmetry breaking.

One can envision circumstances under which T would represent a product of eigenvalues, i.e., a determinant of some transformation, and hence a generalized volume, and that, for example, fundamental symmetry restrictions might well preclude $T = 0$, giving a kind of thermodynamic third law to the system. We explore this below in more detail

Next we examine extend the development to some pathologies affecting real time cognitive systems.

8 No Free Lunch: Inattentional Blindness

The rate tuning theorem analysis of the Appendix permits an inattentional blindness perspective on the famous computational '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 looking at this is to recognize 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 'tuning theorem' argument of the Appendix 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 equivalence classes of problems, although there may well be an optimal algorithm for any given class. The tuning theorem form of the No Free Lunch theorem will apply quite generally to cognitive biological and social structures as well as to massively parallel machines.

Rate distortion, however, occurs when the problem is collapsed into a smaller, simplified, version and then solved. Then there must be a tradeoff between allowed average distortion and the rate of solution: the retina effect. In a very fundamental sense – particularly for real time systems – rate distortion manifolds present a generalization of the converse of the Wolpert/Macready no free lunch arguments. The neural corollary is known as inattentional blindness (Wallace, 2007).

We are led to suggest that there may well be a considerable set of no free lunch-like conundrums confronting highly parallel real-time structures, and that they may interact in distinctly nonlinear ways.

9 Comorbid Pathologies: Developmental Dysfunctions and Critical Periods

Suppose we can operationalize and quantify degrees of both inattentional blindness (IAB) and of overall structure/environment distortion (D) in the actions of a highly parallel cognitive system. The essential assumption is that the (internal) dual information source of a cognitive structure that has low levels of both IAB overfocus and structure/environment distortion will tend to be richer than that of one having greater levels. This is shown in figure 1a, where H is the source uncertainty dual to internal cognitive process, $X = IAB$, and $Y = D$. Regions of low X, Y, near the origin, have greater source uncertainty than those nearby, so $H(X, Y)$ shows a (relatively gentle) peak at the origin, taken here as simply the product of two error functions.

We are, then, focusing on the internal cognitive capacity of the structure itself, as paramatized by degree of overfocus and by the (large scale) distortion between structure implementation and structure impact. That capacity, a purely internal quantity, need not be convex in the parameter D, which is taken to characterize interaction with an external environment, and thus becomes a context for internal measures.

The generalized Onsager argument is shown in figure 1b, where $S = H(X, Y) - XdH/dX - YdH/dY$ is graphed on the Z axis against the $X - Y$ plane, assuming a gentle peak in H at the origin. Peaks in S, according to theory, constitute repulsive system barriers, which must be overcome by external forces. In figure 1b there are three quasi-stable topological resilience modes, in the sense of Wallace (2008a), marked as A, B, and C. The A region is locked in to low levels of both inattentional blindness and distortion, as it sits in a pocket. Forcing the system in either direction, that is, increasing either IAB or D, will, initially, be met by homeostatic attempts to return to the resilience state A, according to this model.

If overall distortion becomes severe in spite of homeostatic mechanisms, the system will then jump to the quasi-stable state B, a second pocket. According to the model, however, once that transition takes place, there will be a tendency for the system to remain in a condition of high distortion. That is, the system will become locked-in to a structure with high distortion in the match between structure implementation and structure impact, but one having lower overall cognitive

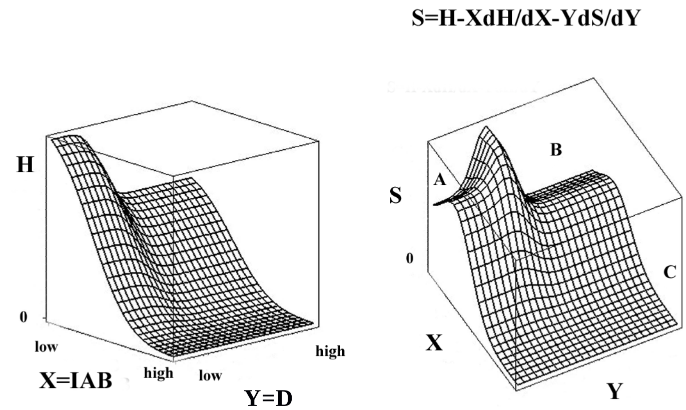

Figure 1: a. Source uncertainty, H, of the dual information source of internal structure cognition, as parametized by degrees of inattentional blindness, $X = IAB$ and distortion $Y = D$ between structure implementation and structure impact. Note the relatively gentle peak at low values of X, Y. Here H is generated as the product of two error functions. b. Generalized Onsager treatment of figure 1a. $S = H(X, Y) - XdH/dX - YdH/dY$. The regions marked A, B, and C represent realms of resilient quasi-stability, divided by barriers defined by the relative peaks in S. Transition among them requires a forcing mechanism. From another perspective, limiting energy or other resources, or imposing stress from the outside – driving down H in figure 1a, would force the system into the lower plain of C, in which the system would then become trapped in states having high levels of distortion and inattentional blindness.

capacity, i.e., a lower value of H in figure 1a.

The third pocket, marked C, is a broad plain in which both IAB and D remain high, a highly overfocused, poorly linked pattern of behavior which will require significant intervention to alter once it reaches such a quasi-stable resilience mode. The structure's cognitive capacity, measured by H in figure 1a, is the lowest of all for this condition of pathological resilience, and attempts to correct the problem – to return to condition A, will be met with very high barriers in S, according to figure 1b. That is, mode C is very highly resilient, although pathologically so, much like the eutrophication of a pure lake by sewage outflow. See Wallace (2008a, b) for a discussion of resilience and literature references.

We can argue that the three quasi-equilibrium configurations of figure 1b represent different dynamical states of the system, and that the possibility of transition between them represents the breaking of the associated symmetry groupoid by external forcing mechanisms. That is, three manifolds representing three different kinds of system dynamics have been patched together to create a more complicated topological structure. For cognitive phenomena, such behavior is likely to be the rule rather than the exception. 'Pure' groupoids are abstractions, and the fundamental questions will involve linkages which break the underlying symmetry.

In all of this, system convergence is not to some fixed state, limit cycle, or pseudorandom strange attractor, but rather to some appropriate set of highly dynamic information sources, i.e., behavior patterns, rather than a fixed 'answer to a computing problem' (Wallace, 2009).

What this model suggests is that sufficiently strong external perturbation can force a highly parallel real-time cognitive structure from a normal, almost homeostatic, developmental path into one involving an analog to a widespread, comorbid, developmental disorder. This is a widely studied pattern for humans and their institutions, reviewed at some length elsewhere (Wallace and Fullilove, 2008; Wallace, 2008b). Indeed, the results of this section might well serve as the foundation of a fairly comprehensive model of chronic developmental dysfunction across a broad class of cognitive systems. One approach might be as follows:

A developmental process can be viewed as involving a sequence of surfaces like figure 1, having, for example, 'critical periods' when the barriers between the normal state A and the pathological states B and C are relatively low. During such a time the system would become highly sensitive to perturbation, and to the onset of a subsequent pathological developmental trajectory. Critical periods might occur during times of rapid learning and/or high system demand for which an energy limitation imposes the need to focus via something like a rate distortion manifold. Cognitive process requires energy through the homologies with free energy density, and more focus at one end necessarily implies less at some other. In a distributed zero sum developmental game, as it were, some cognitive processes must receive more attentional energy than others.

A structure trapped in region C might be said to suffer something much like what Wiegand (2003) describes as the loss of gradient problem, in which one part of a multiple population coevolutionary system comes to dominate the others, creating an impossible situation in which the other participants do not have enough information from which to learn. That is, the cliff just becomes too steep to climb. Wiegand (2003) also characterizes focusing problems in which a two-population coevolutionary process becomes overspecialized on the opponent's weaknesses, effectively a kind of inattentional blindness.

Thus there seems some consonance between our asymptotic analysis of cognitive structural function and current studies of pathologies affecting coevolutionary algorithms (e.g. Ficici et al., 2005; Wallace, 2009). In particular the possibility of historic trajectory, of path dependence, in producing individualized failure modes, is highly disturbing. Under such circumstances there can be no one-size-fits-all amelioration strategy.

10 Topological Constraints

It seems possible to extend this treatment using the methods of section 7, including, then, an *inverse* index of IAB, say $\mathcal{B}$, in the definition of a synergism function, e.g.,

$$T \propto \mathcal{N}^{\alpha}\mathcal{R}^{\beta}R(D)^{\delta}\mathcal{B}^{\gamma},$$

where each component and exponent is real and the underlying components are all positive and nonzero.

More generally, taking T as a product of eigenvalues, we can define it as the determinant of a particular Hessian matrix representing a Morse Function, f, on some underlying, background, manifold, M, characterized in terms of (as yet unspecified) variables $X = (x^1, ..., x^n)$, so that

$$T \propto \det(\mathcal{H}_{i,j}),$$

$$\mathcal{H}_{i,j} \equiv \partial^2 f/\partial x^i \partial x^j.$$

See the Appendix for a brief outline of Morse Theory.

By construction $\mathcal{H}$ has everywhere only nonzero, and indeed, positive, eigenvalues, whose product thereby defines T as a generalized volume. $\mathcal{H}$ becomes, in this model, simply a Jacobean matrix. Thus, and accordingly, all critical points of f have index zero, that is, no eigenvalues of $\mathcal{H}$ are ever negative at any point, and hence at any critical point X_c where $df(X_c) = 0$. This defines a particularly simple topological structure for M: If the interval $[a, b]$ contains a critical value of f with a single critical point x_c, then the topology of the set M_b defined above differs from that of M_a in a manner determined by the index i of the critical point. M_b is then homeomorphic to the manifold obtained from attaching to M_a an i-handle, the direct product of an i-disk and an $(m - i)$-disk. One obtains, in this case, since $i = 0$, the two halves of a sphere with critical points at the top and bottom (Matsumoto, 2002; Pettini, 2007).

The physical natures of $\mathcal{P}, \mathcal{N}, \mathcal{I}, \mathcal{R}, R(D)$, and $\mathcal{B}$ thus impose very stringent constraints on this system, greatly restricting, it appears, possible second order extensions of these statistical models (as in Wallace, 2005, Wallace and Fullilove, 2008). It is as if hierarchical regressions, based on sets of simpler regression models, were themselves to be sharply constrained by inherent structural factors.

11 Discussion and conclusions

A real time system is a composite of a basic computing structure and the embedding environment with which it interacts. The structure, in this treatment, may be a biological system, multiple-workspace distributed cognition social organization, or a computing machine. The structure/environment composite is subject to rate distortion constraints, as are internal processes. These will be driven by entropic torrents and jets that can increase average distortion at all levels of hierarchy. Stabilization, in the face of such degradation force, requires energy and other resources, as implied by the homologies relating information source uncertainty, channel capacity, and free energy density. Cognitive structures are further subject to complicated punctuated phase transition failures depending on average distortion between intent and impact, as well as in the communication between internal components. These are not at all the graceful degradation under stress so hoped for by engineers or managers, and may be greatly compounded by sudden, if predictable, large deviation fluctuations in distortion.

For multiple global workspace systems, for example institutions or high order cognitive machines, punctuated system degradation can be driven by deterioration in communication within or between internal workspaces, or by deterioration in execution.

Further, No Free Lunch constraints similar to inattentional blindness in natural neural processes are likely to become synergistic with distortion torrents and jets to produce complex, path dependent, developmental pathologies that are unlikely to scale well with system size or service load. Such behaviors may make increasingly severe demands on resources to ensure system stabilization. At some point, for a given architecture, a crossover may be reached at which most added resources must be used to simply ensure system stability. Thus it may become necessary to very carefully tune architecture according to both problem type and demand. For biological systems this conundrum has long been recognized as allometric scaling (e.g., White and Seymour, 2005; Speakman, 2005).

Not only will different architectures display different forms of these pathologies, different individual real-time cognitive structures, having particular long-term developmental paths, will be subject to their own special patterns of dysfunction. Remediation under such circumstances seems more akin to an arduous psychiatric intervention than to a simple engineering or medical quick fix. The rate distortion model we have embraced for real time systems sees the structure and its environment as a unit: Environments are always idiosyncratic, and can write images of themselves on developing cognitive structures embedded in them.

Experience suggests that the problems explored here are most likely to be viewed in a manner similar to the discussions of nuclear waste disposal and system safety that took place in the mid 1950's, just as commercial power reactors began to proliferate. Apparently, there must be a certain number of Chernobyl-like failures before such matters gain serious attention.

Indeed, Wallace and Fullilove (2008) and Wallace et al. (2007) explore in some detail how traditions of law and religion have failed to constrain predatory institutional behaviors of distributed cognition leading to more rapid proliferation of HIV infection in the United States, so that, while Chernobyl-like events may be necessary for change, under our socioeconomic and political system, they can fail to be sufficient.

12 Mathematical Appendix

12.1 Groupoids

12.1.1 Basic Ideas

Following Weinstein (1996) closely, a groupoid, G, is defined by a base set A upon which some mapping – a morphism – can be defined. Note that not all possible pairs of states (a_j, a_k) in the base set A can be connected by such a morphism. Those that can define the groupoid element, a morphism $g = (a_j, a_k)$ having the natural inverse $g^{-1} = (a_k, a_j)$. Given such a pairing, it is possible to define 'natural' end-point maps $\alpha(g) = a_j, \beta(g) = a_k$ from the set of morphisms G into A, and a formally associative product in the groupoid $g_1 g_2$ provided $\alpha(g_1 g_2) = \alpha(g_1), \beta(g_1 g_2) = \beta(g_2)$, and $\beta(g_1) = \alpha(g_2)$. Then the product is defined, and associative, $(g_1 g_2)g_3 = g_1(g_2 g_3)$.

In addition, there are natural left and right identity elements λ_g, ρ_g such that $\lambda_g g = g = g\rho_g$ (Weinstein, 1996).

An orbit of the groupoid G over A is an equivalence class for the relation $a_j \sim G a_k$ if and only if there is a groupoid element g with $\alpha(g) = a_j$ and $\beta(g) = a_k$. Following Cannas da Silva and Weinstein (1999), we note that a groupoid is called transitive if it has just one orbit. The transitive groupoids are the building blocks of groupoids in that there is a natural decomposition of the base space of a general groupoid into orbits. Over each orbit there is a transitive groupoid, and the disjoint union of these transitive groupoids is the original groupoid. Conversely, the disjoint union of groupoids is itself a groupoid.

The isotropy group of $a \in X$ consists of those g in G with $\alpha(g) = a = \beta(g)$. These groups prove fundamental to classifying groupoids.

If G is any groupoid over A, the map $(\alpha, \beta) : G \to A \times A$ is a morphism from G to the pair groupoid of A. The image of (α, β) is the orbit equivalence relation $\sim G$, and the functional kernel is the union of the isotropy groups. If $f : X \to Y$ is a function, then the kernel of f, $ker(f) = [(x_1, x_2) \in X \times X : f(x_1) = f(x_2)]$ defines an equivalence relation.

Groupoids may have additional structure. As Weinstein (1996) explains, a groupoid G is a topological groupoid over a base space X if G and X are topological spaces and α, β and multiplication are continuous maps. A criticism sometimes applied to groupoid theory is that their classification up to isomorphism is nothing other than the classification of equivalence relations via the orbit equivalence relation and groups via the isotropy groups. The imposition of a compatible topological structure produces a nontrivial interaction between the two structures. Below we will introduce a metric structure on manifolds of related information sources, producing such interaction.

In essence, a groupoid is a category in which all morphisms have an inverse, here defined in terms of connection to a base point by a meaningful path of an information source dual to a cognitive process.

As Weinstein (1996) points out, the morphism (α, β) suggests another way of looking at groupoids. A groupoid over A identifies not only which elements of A are equivalent to one another (isomorphic), but *it also parametizes the different ways (isomorphisms) in which two elements can be equivalent*, i.e., all possible information sources dual to some cognitive process. Given the information theoretic characterization of cognition presented above, this produces a full modular cognitive network in a highly natural manner.

Brown (1987) describes the fundamental structure as follows:

> A groupoid should be thought of as a group with many objects, or with many identities... A groupoid with one object is essentially just a group. So the notion of groupoid is an extension of that of groups. It gives an additional convenience, flexibility and range of applications...
>
> EXAMPLE 1. A disjoint union [of groups] $G = \cup_\lambda G_\lambda, \lambda \in \Lambda$, is a groupoid: the product ab is defined if and only if a, b belong to the same G_λ, and ab is then just the product in the group G_λ. There is an identity 1_λ for each $\lambda \in \Lambda$. The maps α, β coincide and map G_λ to λ, $\lambda \in \Lambda$.
>
> EXAMPLE 2. An equivalence relation R on [a set] X becomes a groupoid with $\alpha, \beta : R \to X$ the two projections, and product $(x, y)(y, z) = (x, z)$ whenever $(x, y), (y, z) \in R$. There is an identity, namely (x, x), for each $x \in X$...

Weinstein (1996) makes the following fundamental point:

> Almost every interesting equivalence relation on a space B arises in a natural way as the orbit equivalence relation of some groupoid G over B. Instead of dealing directly with the orbit space B/G as an object in the category S_{map} of sets and mappings, one should consider instead the groupoid G itself as an object in the category G_{htp} of groupoids and homotopy classes of morphisms.

The groupoid approach has become quite popular in the study of networks of coupled dynamical systems which can be defined by differential equation models, (e.g., Golubitsky and Stewart 2006).

12.1.2 Global and Local Symmetry Groupoids

Here we follow Weinstein (1996) fairly closely, using his example of a finite tiling.

Consider a tiling of the euclidean plane R^2 by identical 2 by 1 rectangles, specified by the set X (one dimensional) where the grout between tiles is $X = H \cup V$, having $H = R \times Z$ and $V = 2Z \times R$, where R is the set of real numbers and Z the integers. Call each connected component of $R^2 \backslash X$, that is, the complement of the two dimensional real plane intersecting X, a tile.

Let Γ be the group of those rigid motions of R^2 which leave X invariant, i.e., the normal subgroup of translations by elements of the lattice $\Lambda = H \cap V = 2Z \times Z$ (corresponding to corner points of the tiles), together with reflections through each of the points $1/2\Lambda = Z \times 1/2Z$, and across the horizontal and vertical lines through those points. As noted by Weinstein (1996), much is lost in this coarse-graining, in particular the same symmetry group would arise if we replaced X entirely by the lattice Λ of corner points. Γ retains no information about the local structure of the tiled plane. In the case of a real tiling, restricted to the finite set $B = [0, 2m] \times [0, n]$ the symmetry group shrinks drastically: The subgroup leaving $X \cap B$ invariant contains just four elements even though a repetitive pattern is clearly visible. A two-stage groupoid approach recovers the lost structure.

We define the transformation groupoid of the action of Γ on R^2 to be the set

$$G(\Gamma, R^2) = \{(x, \gamma, y | x \in R^2, y \in R^2, \gamma \in \Gamma, x = \gamma y\},$$

with the partially defined binary operation

$$(x, \gamma, y)(y, \nu, z) = (x, \gamma\nu, z).$$

Here $\alpha(x, \gamma, y) = x$, and $\beta(x, \gamma, y) = y$, and the inverses are natural.

We can form the restriction of G to B (or any other subset of R^2) by defining

$$G(\Gamma, R^2)|_B = \{g \in G(\Gamma, R^2) | \alpha(g), \beta(g) \in B\}$$

[1]. An orbit of the groupoid G over B is an equivalence class for the relation

$x \sim_G y$ if and only if there is a groupoid element g with $\alpha(g) = x$ and $\beta(g) = y$.

Two points are in the same orbit if they are similarly placed within their tiles or within the grout pattern.

[2]. The isotropy group of $x \in B$ consists of those g in G with $\alpha(g) = x = \beta(g)$. It is trivial for every point except those in $1/2\Lambda \cap B$, for which it is $Z_2 \times Z_2$, the direct product of integers modulo two with itself.

By contrast, embedding the tiled structure within a larger context permits definition of a much richer structure, i.e., the identification of local symmetries.

We construct a second groupoid as follows. Consider the plane R^2 as being decomposed as the disjoint union of $P_1 = B \cap X$ (the grout), $P_2 = B \backslash P_1$ (the complement of P_1 in B, which is the tiles), and $P_3 = R^2 \backslash B$ (the exterior of the tiled room). Let E be the group of all euclidean motions of the plane, and define the local symmetry groupoid G_{loc} as the set of triples (x, γ, y) in $B \times E \times B$ for which $x = \gamma y$, and for which y has a neighborhood $\mathcal{U}$ in R^2 such that $\gamma(\mathcal{U} \cap P_i) \subseteq P_i$ for $i = 1, 2, 3$. The composition is given by the same formula as for $G(\Gamma, R^2)$.

For this groupoid-in-context there are only a finite number of orbits:

$\mathcal{O}_1$ = interior points of the tiles.

$\mathcal{O}_2$ = interior edges of the tiles.

$\mathcal{O}_3$ = interior crossing points of the grout.

$\mathcal{O}_4$ = exterior boundary edge points of the tile grout.

$\mathcal{O}_5$ = boundary 'T' points.

$\mathcal{O}_6$ = boundary corner points.

The isotropy group structure is, however, now very rich indeed:

The isotropy group of a point in $\mathcal{O}_1$ is now isomorphic to the entire rotation group O_2.

It is $Z_2 \times Z_2$ for $\mathcal{O}_2$.

For $\mathcal{O}_3$ it is the eight-element dihedral group D_4.

For $\mathcal{O}_4, \mathcal{O}_5$ and $\mathcal{O}_6$ it is simply Z_2.

These are the 'local symmetries' of the tile-in-context.

12.2 Generalized Onsager Theory

Understanding the time dynamics of groupoid-driven information systems away from the kind of phase transition critical points described above requires a phenomenology similar to the Onsager relations of nonequilibrium thermodynamics. This also leads to a general theory involving large-scale topological changes in the sense of Morse theory.

If the Groupoid Free Energy of a biological process is parametized by some vector of quantities $\mathbf{K} \equiv (K_1, ..., K_m)$, then, in analogy with nonequilibrium thermodynamics, gradients in the K_j of the *disorder*, defined as

$$S_G \equiv F_G(\mathbf{K}) - \sum_{j=1}^{m} K_j \partial F_G / \partial K_j$$

(28)

become of central interest.

Equation (28) is similar to the definition of entropy in terms of the free energy of a physical system.

Pursuing the homology further, the generalized Onsager relations defining temporal dynamics of systems having a GFE become

$$dK_j/dt = \sum_i L_{j,i} \partial S_G / \partial K_i,$$

(29)

where the $L_{j,i}$ are, in first order, constants reflecting the nature of the underlying cognitive phenomena. The L-matrix is to be viewed empirically, in the same spirit as the slope and intercept of a regression model, and may have structure far different than familiar from more simple chemical or physical processes. The $\partial S_G/\partial K$ are analogous to thermodynamic forces in a chemical system, and may be subject to override by external physiological or other driving mechanisms: biological and cognitive phenomena, unlike simple physical systems, can make choices as to resource allocation.

That is, an essential contrast with simple physical systems driven by (say) entropy maximization is that complex biological or cognitive structures can make decisions about resource allocation, to the extent resources are available. Thus resource availability is a context, not a determinant, of behavior.

Equations (28) and (29) can be derived in a simple parameter-free covariant manner which relies on the underlying topology of the information source space implicit to the development (e.g., Wallace and Wallace, 2008b). We will not pursue that development here.

The dynamics, as we have presented them so far, have been noiseless, while biological systems are always very noisy. Equation (29) might be rewritten as

$$dK_j/dt = \sum_i L_{j,i} \partial S_G / \partial K_i + \sigma W(t)$$

where σ is a constant and $W(t)$ represents white noise. This leads directly to a family of classic stochastic differential equations having the form

$$dK_t^j = L^j(t, \mathbf{K})dt + \sigma^j(t, \mathbf{K})dB_t,$$

(30)

where the L^j and σ^j are appropriately regular functions of t and $\mathbf{K}$, and dB_t represents the noise structure, and we have readjusted the indices.

Further progress in this direction requires introduction of methods from stochastic differential geometry and related topics in the sense of Emery (1989). The obvious inference is that noise – not necessarily 'white' – can serve as a tool to shift the system between various topological modes, as a kind of crosstalk and the source of a generalized stochastic resonance.

Effectively, topological shifts between and within dynamic manifolds constitute another theory of phase transitions (Pettini, 2007),and this phenomenological Onsager treatment would likely be much enriched by explicit adoption of a Morse theory perspective.

12.3 The Tuning Theorem

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

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

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

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

(31)

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

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

(32)

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

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

(33)

See, for example, Ash (1990), Khinchin (1957) or Cover and Thomas (1991) 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

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

(34)

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, Ash (1990), Khinchin (1957) and Cover and Thomas (1991) provide details.

This approach can be, in a sense, inverted to give a tuning theorem which parsimoniously describes the essence of the Rate Distortion Manifold.

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 other signaling 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

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

(35)

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, a tunable retina-like Rate Distortion Manifold. It seems likely, however, than for many important systems $P(X)$ and $P(Y)$ will interpenetrate, to use Richard Levins' terminology. That is, $P(X)$ and $P(Y)$ will affect each other in characteristic ways, so that some form of mutual tuning may be the most effective strategy.

12.4 Morse Theory

Morse theory examines relations between analytic behavior of a function – the location and character of its critical points – and the underlying topology of the manifold on which the function is defined. We are interested in a number of such functions, for example information source uncertainty on a parameter space and 'second order' iterations involving parameter manifolds determining critical behavior, for example sudden onset of a giant component in the mean number model (Wallace and Wallace, 2008), and universality class tuning in the mean field model of the next section. These can be reformulated from a Morse theory perspective. Here we follow closely the elegant treatments of Pettini (2007) and Kastner (2006).

The essential idea of Morse theory is to examine an n-dimensional manifold M as decomposed into level sets of some function $f : M \to \mathbf{R}$ where $\mathbf{R}$ is the set of real numbers. The a-level set of f is defined as

$$f^{-1}(a) = \{x \in M : f(x) = a\},$$

the set of all points in M with $f(x) = a$. If M is compact, then the whole manifold can be decomposed into such slices in a canonical fashion between two limits, defined by the minimum and maximum of f on M. Let the part of M below a be defined as

$$M_a = f^{-1}(-\infty, a] = \{x \in M : f(x) \leq a\}.$$

These sets describe the whole manifold as a varies between the minimum and maximum of f.

Morse functions are defined as a particular set of smooth functions $f : M \to \mathbf{R}$ as follows. Suppose a function f has a critical point x_c, so that the derivative $df(x_c) = 0$, with critical value $f(x_c)$. Then f is a Morse function if its critical points are nondegenerate in the sense that the Hessian matrix of second derivatives at x_c, whose elements, in terms of local coordinates are

$$\mathcal{H}_{i,j} = \partial^2 f/\partial x^i \partial x^j,$$

has rank n, which means that it has only nonzero eigenvalues, so that there are no lines or surfaces of critical points and, ultimately, critical points are isolated.

The index of the critical point is the number of negative eigenvalues of $\mathcal{H}$ at x_c.

A level set $f^{-1}(a)$ of f is called a critical level if a is a critical value of f, that is, if there is at least one critical point $x_c \in f^{-1}(a)$.

Again following Pettini (2007), the essential results of Morse theory are:

[1] If an interval $[a, b]$ contains no critical values of f, then the topology of $f^{-1}[a, v]$ does not change for any $v \in (a, b]$. Importantly, the result is valid even if f is not a Morse function, but only a smooth function.

[2] If the interval $[a, b]$ contains critical values, the topology of $f^{-1}[a, v]$ changes in a manner determined by the properties of the matrix H at the critical points.

[3] If $f : M \to \mathbf{R}$ is a Morse function, the set of all the critical points of f is a discrete subset of M, i.e., critical points are isolated. This is Sard's Theorem.

[4] If $f : M \to \mathbf{R}$ is a Morse function, with M compact, then on a finite interval $[a, b] \subset \mathbf{R}$, there is only a finite number of critical points p of f such that $f(p) \in [a, b]$. The set of critical values of f is a discrete set of $\mathbf{R}$.

[5] For any differentiable manifold M, the set of Morse functions on M is an open dense set in the set of real functions of M of differentiability class r for $0 \leq r \leq \infty$.

[6] Some topological invariants of M, that is, quantities that are the same for all the manifolds that have the same topology as M, can be estimated and sometimes computed exactly once all the critical points of f are known: Let the Morse numbers $\mu_i (i = 0, ..., m)$ of a function f on M be the number of critical points of f of index i, (the number of negative eigenvalues of H). The Euler characteristic of the complicated manifold M can be expressed as the alternating sum of the Morse numbers of any Morse function on M,

$$\chi = \sum_{i=1}^{m} (-1)^i \mu_i.$$

The Euler characteristic reduces, in the case of a simple polyhedron, to

$$\chi = V - E + F$$

where V, E, and F are the numbers of vertices, edges, and faces in the polyhedron.

[7] Another important theorem states that, if the interval $[a, b]$ contains a critical value of f with a single critical point x_c, then the topology of the set M_b defined above differs from that of M_a in a way which is determined by the index, i, of the critical point. Then M_b is homeomorphic to the manifold obtained from attaching to M_a an i-handle, i.e., the direct product of an i-disk and an $(m - i)$-disk.

Again, Pettini (2007) contains both mathematical details and further references. See, for example, Matusmoto (2002) or the classic by Milnor (1963).

13 References

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

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

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

Bos, R., 2007, Continuous representations of groupoids, arXiv:math/0612639.

Buneci, M., 2003, *Representare de Groupoizi*, Editura Mirton, Timisoara.

Cannas Da Silva, A., and A. Weinstein, 1999, *Geometric Models for Noncommutative Algebras*, American Mathematical Society, Providence, RI.

Cover, T., and J. Thomas, 1991, *Elements of Information Theory*, John Wiley and Sons, New York.

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

Dretske, F., 1981, *Knowledge and the flow of information*, MIT Press, Cambridge, MA.

Dretske, F., 1988, *Explaining behavior*, MIT Press, Cambridge, MA.

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

Emery, M., 1989, *Stochastic Calculus on Manifolds*, Universititext Series, Springer, 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*, L. Fogel, P. Angeline, and T. Back, eds., pp. 163-169, MIT Press, Cambridge, MA.

Feller, W., 1971, *An Introduction to Probability Theory and its Applications*, John Wiley and Sons, New York.

Ficici, S., O. Milnik, and J. Pollack, 2005, A game-theoretic and dynamical systems analysis of selection methods in coevolution, *IEEE Transactions on Evolutionary Computation*, 9:580-602.

Fredlin M., and A. Wentzell, 1998, *Random Perturbations of Dynamical Systems*, Springer, New York.

Golubitsky, M., and I. Stewart, 2006, Nonlinear dynamics and networks: the groupoid formalism, *Bulletin of the American Mathematical Society*, 43:305-364.

Granovetter, M., 1973, The strength of weak ties, *American Journal of Sociology*, 78:1360-1380.

Landau, L., and E. Lifshitz, 2007, *Statistical Physics, 3rd Edition*, Part I, Elsevier, New York.

Luchinsky, D., 1997, On the nature of large fluctuations in equilibrium systems: observations of an optimal force, *Journal of Physics A*, 30:L577-L583.

Matsumoto, Y., 2002, *An Introduction to Morse Theory*, American Mathematical Society, Providence, RI.

Onsager, L., and S. Machlup, 1953, Fluctuations and irreversible processes, *Physical Review*, 91:1505-1512.

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

Protter, P., 1995, *Stochastic Integration and Differential Equations: A New Approach*, Springer, New York.

Skierski, M., A. Grundland, and J. Tuszynski, 1989, Analysis of the three-dimensional time-dependent Landau-Ginzburg equation and its solutions, *Journal of Physics A*(Math. Gen.), 22:3789-3808.

Speakman. J., Body size, energy metabolism, and lifespan, *Journal of Experimental Biology*, 208(Pt. 9):1717-1730.

Wallace, R., 2005, *Consciousness: A mathematical treatment of the Global Neuronal Workspace model*, Springer, New York.

Wallace R., 2006, Pitfalls in biological computing: canonical and idiosyncratic dysfunction of conscious machines, *Mind and Matter*, 4:91-113.

Wallace R., 2007, Culture and inattentional blindness, *Journal of Theoretical Biology*, 245:378-390.

Wallace, R., M. Fullilove, R. Fullilove, and D. Wallace, 2007, Collective consciousness and its pathologies: Understanding the failure of AIDS control and treatment in the United States, *Theoretical Biology and Medical Modelling*, 4:10.

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

Wallace, R., 2008b, Developmental disorders as pathological resilience domains, *Ecology and Society* 13(1):29.

Wallace, R., 2009, Toward formal models of biologically inspired, highly parallel machine cognition: II programming coevolutionary machines. Submitted.

Wallace R., and M. Fullilove, 2008, *Collective Consciousness and Its Discontents: Institutional Distributed Cognition, Racial Policy, and Public Health in the United States*, Springer, New York.

Wallace R., and D. Wallace, 2008, Punctuated equilibrium in statistical models of generalized coevolutionary resilience: how sudden ecosystem transitions can entrain both phenotype expression and Darwinian selection, *Transactions on Computational Systems Biology IX*, LNBI 5121:23-85.

Weinstein, A., 1996, Groupoids: unifying internal and external symmetry, *Notices of the American Mathematical Society*, 43:744-752.

White, C., and R. Seymour, 2005, Allometric scaling of mammalian metabolism, *Journal of Experimental Biology*, 208(Pt. 9):1611-1619.

Wiegand, R., 2003, *An analysis of cooperative coevolutionary algorithms*, PhD Thesis, George Mason University.

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

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