Version 3

Entering the blackboard jungle: canonical dysfunction in conscious machines

Rodrick Wallace, Ph.D.
The New York State Psychiatric Institute[*]

October 31, 2005

Abstract

The central paradigm of Artificial Intelligence is rapidly shifting toward biological models for both robotic devices and systems performing such critical tasks as network management and process control. Here we apply recent mathematical analysis of the necessary conditions for consciousness in humans in an attempt to gain some understanding of likely canonical failure modes inherent to a broad class of global workspace/blackboard machines designed to emulate biological functions. Similar problems are likely to confront other possible architectures, although their mathematical description may be far less straightforward.

Key words: artificial intelligence, autonomic computing, cancer, cognition, consciousness, information theory, mental disorder

Introduction

[*] Address correspondence to: R. Wallace, PISCS Inc., 549 W. 123 St., New York, NY, 10027 USA. Telephone (212) 865-4766, email rdwall@ix.netcom.com. Affiliation is for identification only.

The artificial intelligence community, after half a century of furious work, has at last come to understand the serious limitations of prevailing approaches. The preamble to the forthcoming "50th Anniversary Summit of Artificial Intelligence", (Summit, 2005), which includes many of the leading lights of AI, states that

> "Despite its advances in the last 50 years [since a defining meeting on AI held at Dartmouth College in 1956], it is clear that the original goals set by the first generation of AI visionaries have not been reached... [T]he current landscape of research reveals how little we know about how biological brains achieve their remarkable functionalitiy... [W]e do not understand the cultural and social processes that have helped to shape human intelligence... What has been missed is - we believe - how important embodiment and the interaction with the world are as the basis for thinking... [An important objective of the forthcoming meeting] will be to do away with the computational metaphor that has been haunting AI for 50 years: the brain is not a computer! A merely computational approach to understanding natural intelligence and realizing artificial forms of it (which was proposed in 1956 [at the Dartmouth meeting]) has been shown to be inadequate. It is time to move beyond computation. Recently, artificial intelligence researchers have become aware of the fact that uncovering the mechanisms underlying the dynamics of the mutual and reciprocal interaction of body, brain, and environment is of crucial importance. We believe that it is important to reassess the field based on this recent paradigm change..."

It appears that the organizers of the new AI 'Summit', while rightly questioning current paradigms, and moving to adopt a more biologically-oriented approach, do not quite understand the implications of recent empirical work in support of Bernard Baars' Global Workspace (GW) model of consciousness (e.g. Baars, 1988; Baars and Franklin, 2003; Dehaene and Naccache, 2001), which, although frequently reexpressed or reinvented by various researchers at various times (esp. Newell, 1990), implicitly underlies much of their discussion. Baars' model, over the past two decades, has received increasing experimental verification (e.g. Massimini et al, 2005). Since it particularly attempts to properly represent the matter of embedding and interpenetrating contexts, it provides a basis for a machine architecture likely to emulate or surpass natural intelligence in more than just playing variants of chess.

My own work provides a rigorous mathematical formulation of the GW blackboard model, in terms of an iterated, second-order, contextually-embedded, hierarchical General Cognitive Model (GCM) crudely analogous to hierarchical regression. It is, however, based on the Shannon-McMillan rather than on the Central Limit Theorem, and is strongly supplemented by methodologies from topological manifold theory and differential geometry (Wallace, 2005a, b, c). Recent results (Wallace, 2005c) suggest that, in fact, it should be possible to make a rigorous theory of 'all possible' GW blackboard models. This would take the role that the Church lambda calculus plays for conventional machine architecture, particularly when iterated to second order. It should be possible to implement machines which instantiate these models in silicon or software, and compare and contrast designed machines with others reflecting empirical data on the observed properties of human or animal consciousness and cognition. Such contrast and comparison would greatly deepen our understanding of biological consciousness.

This being said, the promise of conscious machines which interact closely with their embedding systems and external environments seems indeed vast: robots which can do the functional equivalent of riding a bicycle in heavy traffic; intuitive network managers which can sense oncoming difficulties and adapt prior to their occurrence; conscious search engines and marketing machines whose efficiency is not constrained by the path-dependent forms which evolutionary history has enforced for humans or animals, and so on. AI, it seems, is finally going to get it right.

A principal impetus, at least in the United States, for seeking new models for software and system design, what IBM calls its 'autonomic computing initiative', is the increasing difficulty of maintaining current software, with more and more corporate resources progressively dedicated to 'bug fixes', recovering from, and insulating against, crashes, security failures, and the like. The thought has been that autonomic computing would permit programs and systems to, in large measure, heal or protect themselves, with minimal human intervention. Some time invested in creating such an environment, it is felt, would pay huge dividends down the road in decreased maintenance costs. The initiative seems clearly designed to produce a first order cognitive, but not a second order conscious, machine. IBM too, it seems, is finally going to get it right. Can Microsoft be far behind?

One is, much like the proverbial small child at the back of the room, driven to ask an obvious question: What are the likely canonical and idiosyncratic failure modes of the coming wave of second order global workspace blackboard machines? Under what circumstances can such machines be used without great risk in mission-critical tasks, for example driving a car or truck, managing a nuclear power plant, a chemical factory, a national power or communications grid, or even a system of ATM machines?

A formal exploration of the global workspace model suggests something of the pitfalls facing conscious or cognitive machines and their designers. We begin with a simplified analysis focusing on modular networks of interacting cognitive substructures, and particularly study the importance of their embedding in progressively larger systems. More complicated examples, involving renormalization treatment of phase transitions affecting information sources, iterated to second order, can be found in Wallace (2005a).

The simplest modular network blackboard model

Cognition as 'language' Cognition is not consciousness. Indeed, most mental, and many physiological, functions, while cognitive in a particular formal sense, hardly ever become entrained into the Global Workspace of consciousness. For example, one seldom is able to consciously regulate immune function, blood pressure, or the details of binocular tracking and bipedal motion, except to decide 'what shall I look at', 'where shall I walk'. Nonetheless, many cognitive processes, conscious or unconscious, appear intimately related to 'language', broadly speaking. The construction is surprisingly straightforward (Wallace, 2000, 2005a).

Atlan and Cohen (1998) and Cohen (2000) argue, in the context of immune cognition, that the essence of cognitive function involves comparison of a perceived signal with an internal, learned picture of the world, and then, upon that comparison, choice of one response from a much larger repertoire of possible responses.

Cognitive pattern recognition-and-response, from this view, proceeds by functionally combining an incoming external sensory signal with an internal ongoing activity – incorporating the learned picture of the world – and triggering an appropriate action based on a decision that the pattern of sensory activity requires a response.

More formally, a pattern of sensory input is mixed in an unspecified but systematic 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 algorithmic composition of internal and external signals.

This path is fed into a highly nonlinear, but otherwise similarly unspecified, nonlinear decision oscillator 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 interest are paths x which trigger pattern recognition-and-response exactly once. 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 exactly once to the event $h(x) \in B_1$. Thus $h(a_0, ..., a_j) \in B_0$ for all $j < m$, but $h(a_0, ..., a_m) \in B_1$. Wallace (2005a) examines the possibility of more complicated schemes as well.

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 having $h(a_0) \in B_0$ 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 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.

We 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 is possible (Wallace, 2005a).

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 (Wallace, 2005a).

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 Khinchine (1957), Ash (1990), or Cover and Thomas (1991) for the standard details.

The giant component A formal equivalence class algebra (and hence a groupoid, sensu Weinstein, 1996) can be constructed by choosing different origin points a_0 and defining equivalence by the existence of a high probability meaningful path connecting two points. Disjoint partition by equivalence class, analogous to orbit equivalence classes for dynamical systems, defines the vertices of the proposed network of cognitive dual languages. Each vertex then represents a different information source dual to a cognitive process.

We now suppose that linkages can fleetingly occur between the ordinarily disjoint cognitive modules defined by this algebra. In the spirit of Wallace

(2005a), this is represented by establishment of a non-zero mutual information measure between them: cross-talk.

Wallace (2005a) describes this structure in terms of fixed magnitude disjunctive strong ties which give the equivalence class partitioning of modules, and nondisjunctive weak ties which link modules across the partition, and parametizes the overall structure by the average strength of the weak ties, to use Granovetter's (1973) term. By contrast the approach here, initially, is to simply look at the average number of fixed-strength nondisjunctive links in a random topology. These are obviously the two analytically tractable limits of a much more complicated regime which we believe ultimately includes 'all possible' global workspace models.

Since we know nothing about how the cross-talk connections can occur, we will – for purposes of illustration only – assume they are random and construct a random graph in the classic Erdos/Renyi manner. Suppose there are M disjoint cognitive modules – M elements of the equivalence class algebra of languages dual to some cognitive process – which we now take to be the vertices of a possible graph.

As Corless et al. (1996) discuss, when a graph with M vertices has $m = (1/2)aM$ edges chosen at random, for $a > 1$ it almost surely has a giant connected component having approximately gM vertices, with

$$g(a) = 1 + W(-a\exp(-a))/a,$$

(2)

where W is the Lambert-W function defined implicitly by the relation

$$W(x)\exp(W(x)) = x.$$

(3)

Figure 1 shows $g(a)$, displaying what is clearly a sharp phase transition at $a = 1$.

Such a phase transition initiates a new, collective, shifting, cognitive phenomenon: the Global Workspace, a tunable blackboard defined by a set of cross-talk mutual information measures between interacting unconscious cognitive submodules. The source uncertainty, H, of the language dual to the collective cognitive process, which defines the richness of the cognitive language of the workspace, will grow as some function of g, as more and more unconscious processes are incorporated into it. Wallace (2005a) examines what, in effect, are the functional forms $H \propto \exp(\alpha g), \alpha \ln[1/(1-g)]$, and $(1/(1-g))^{\delta}$, letting $R = 1/1 - g$ define a 'characteristic length' in the renormalization scheme. While these all have explicit solutions for the renormalization calculation (mostly in terms of the Lambert-W function), other, less tractable, expressions are certainly plausible, for example $H \propto g^{\gamma}, \gamma > 0$, γ real.

Given a particular $H(g)$, the quite different approach of Wallace (2005a) involves adjusting universality class parameters of the phase transition, a matter requiring much mathematical development.

By contrast, in this new class of models, the degree of clustering of the graph of cognitive modules might, itself, be tunable, producing a variable threshold for consciousness: a topological shift, which should be observable from brain-imaging studies. Second order iteration would lead to an analog of the hierarchical cognitive model of Wallace (2005a).

Wallace (2005a) focuses on changing the average strength of weak ties between unconscious submodules rather than the average number of fixed-strength weak ties as is done here, and tunes the universality class exponents of the phase transition, which may also imply subtle shifts in underlying topology.

Following Albert and Barabasi (2002, Section V), we note that real networks differ from random graphs in that their degree distribution, the probability of k linkages between vertices, often follows a power law $P(k) \approx k^{-\gamma}$ rather than the Poisson distribution of random networks,

$P(k) = a^k \exp(-a)/k!, k \geq 0$. Since power law networks do not have any characteristic scale, they consequently termed scale-free.

It is possible to extend the Erdos/Renyi threshold results to such 'semi-random' graphs. For example, Luczak (1992) has shown that almost all random graphs with a fixed degree smaller than 2 have a unique giant cluster. Molloy and Reed (1995, 1998) proved that, for a random graph with degree distribution $P(k)$, an infinite cluster emerges almost surely when

$$Q \equiv \sum_{k\geq 1} k(k-2)P(k) > 0.$$

(4)

Following Volz, (2004), cluster tuning of random networks leads to a counterintuitive result. Define the clustering coefficient C as the proportion of triads in a network out of the total number of potential triads, i.e.

$$C = \frac{3N_\Delta}{N_3},$$

(5)

where N_Δ is the number of triads in the network and N_3 is the number of connected triples of nodes, noting that in every triad there are three connected nodes. Taking the approach of Molloy and Reed (1995), Volz shows quite directly that, for a random network with parameter a, at cluster value C, there is a critical value given by

$$a_C = \frac{1}{1-C-C^2}.$$

(6)

If $C = 0$, i.e. no clustering, then the giant component forms when $a = 1$. Increasing C *raises* the average number of edges which must be present for a giant component to form. For $C \geq \sqrt{5}/2 - 1/2$, which is precisely the Golden Section, where the denominator in this expression vanishes, no giant component can form, regardless of a. Not all network topologies, then, can actually support a giant component, and hence, in this model, consciousness. This is of some importance, having obvious and deep implications ranging from the evolutionary history of consciousness to the nature of sleep.

A more complete exploration of the giant component can be found, e.g. in Newman et al. (2001), especially the discussion leading to their figure 4. In general, 'tuning' of the GC will generate a family of curves similar to figure 1, but with those having threshold to the right of that in the plot 'topping out' at limits progressively less than 1: higher thresholds seem usually to imply smaller giant components. In sum, the giant component is itself highly tunable, replicating, in this model, the fundamental stream of consciousness.

Note that we do not, in this paper, address the essential matter of how the system of interacting cognitive modules behaves away from critical points, particularly in the presence of 'external gradients'. Answering this question requires the imposition of generalized Onsager relations, which introduce complications of topological 'rate distortion manifolds', metric structures, and the like (e.g. Wallace, 2005a, b).

Mutual and reciprocal interaction: evading the mereological fallacy Just as a higher order information source, associated with the GC of a random or semirandom graph, can be constructed out of the interlinking of unconscious cognitive modules by mutual information, so too external information sources, for example in humans the cognitive immune and other physiological systems, and embedding sociocultural structures, can be represented as slower-acting information sources whose influence on the GC can be felt in a collective mutual information measure. For machines these would be the onion-like 'structured environment', to be viewed as among Baars' contexts (Baars, 1988; Baars and Franklin, 2003). The collective mutual information measure will, through the Joint Asymptotic Equipartition Theorem which generalizes the Shannon-McMillan Theorem, be the splitting criterion for high and low probability joint paths across the entire system.

The tool for this is network information theory (Cover and Thomas, 1991, p. 387). Given three interacting information sources, Y_1, Y_2, Z, the splitting criterion, taking Z as the 'external context', is given by

$$I(Y_1, Y_2|Z) = H(Z) + H(Y_1|Z) - H(Y_1, Y_2, Z),$$

(7)

where $H(..|..)$ and $H(.., .., ..)$ represent conditional and joint uncertainties (Khinchine, 1957; Ash, 1990; Cover and Thomas, 1991).

This generalizes to

$$I(Y_1, ...Y_n|Z) = H(Z) + \sum_{j=1}^{n} H(Y_j|Z) - H(Y_1, ..., Y_n, Z).$$

(8)

If we assume the Global Workspace/GC/blackboard to involve a very rapidly shifting, and indeed highly tunable, dual information source X, embedding contextual cognitive modules like the immune system will have a set of significantly slower-responding sources $Y_j, j = 1..m$, and external social, cultural and other 'environmental' processes will be characterized by even more slowly-acting sources $Z_k, k = 1..n$. Mathematical induction on equation (8) gives a complicated expression for a mutual information splitting criterion which we write as

$$I(X|Y_1, .., Y_m|Z_1, .., Z_n).$$

(9)

This encompasses a fully interpenetrating 'biopsychosociocultural' structure for individual human or machine consciousness, one in which Baars' contexts act as important, but flexible, boundary conditions, defining the underlying topology available to the far more rapidly shifting global workspace (Wallace, 2005a, b).

This result does not commit the mereological fallacy which Bennett and Hacker (2003) impute to excessively neurocentric perspectives on consciousness in humans, that is, the mistake of imputing to a part of a system the characteristics which require functional entirety. The underlying concept of this fallacy should extend to machines interacting with their environments, and its baleful influence probably accounts for a significant part of AI's failure to deliver. See Wallace (2005a) for further discussion.

Punctuation phenomena

As quite a number of researchers have noted, in one way or another, – see Wallace, (2005a) for discussion – equation (1),

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

is homologous to the thermodynamic limit in the definition of the free energy density of a physical system. This has the form

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

(10)

where F is the free energy density, K the inverse temperature, V the system volume, and $Z(K)$ is the partition function defined by the system hamiltonian.

Wallace (2005a) shows at some length how this homology permits the natural transfer of renormalization methods from statistical mechanics to information theory. In the spirit of the Large Deviations Program of applied probability theory, this produces phase transitions and analogs to evolutionary punctuation in systems characterized by piecewise, adiabatically stationary, ergodic information sources. These 'biological' phase changes appear to be ubiquitous in natural systems and can be expected to dominate machine behaviors as well, particularly those which seek to emulate biological paradigms. Wallace (2002) uses these arguments to explore the differences and similiarities between evolutionary punctuation in genetic and learning plateaus in neural systems. Punctuated phenomena will emerge as important in the discussions below of subtle information machine malfunctions.

The dysfunctions of consciousness and intelligence

Somewhat surprisingly, equation (9), informed by the homology with equation (10), permits general discussion of the failure modes of global workspace blackboard machines, in particular of their second order iteration which appears to be the analog to consciousness in humans and other higher animals.

The foundation for this lies in the Rate Distortion Theorem. Under the conditions of that theorem, equation (9) is the splitting criterion defining the maximum rate at which an external information source can write an image of itself having a given maximum of distortion, according to some defined measure (Cover and Thomas, 1991; Dembo and Zeitouni, 1998). Inverting the argument, equation (9) suggests that an external information source can, if given enough time, write an image of itself upon consciousness. If that external source is pathogenic in terms of machine structure, then, given sufficient exposure, some measure of consciousness dysfunction becomes inevitable.

Comorbid mind/body disorders in humans are worth exploring (Wallace, 2004).

Mental disorders in humans are not well understood (e.g. Wallace, 2005a, Ch. 6). Indeed, such classifications as the *Diagnostic and Statistical Manual of Mental Disorders - fourth edition*, (DSM-IV, 1994), the standard descriptive nosology in the US, have been characterized as 'prescientific' by Gilbert (2001) and others. Arguments from genetic determinism fail, in part because of an apparent genetic bottleneck which, early in our species' history, resulted in an overall genetic diversity less than that observed within and between contemporary chimpanzee populations. Arguments from psychosocial stress fare better, but are affected by the apparently complex and contingent developmental paths determining the onset of schizophrenia, one of the most prevalent serious mental disorders, dementias, psychoses, and so forth, some of which may be triggered in utero by exposure to infection, low birthweight, or other stressors. Gilbert suggests an evolutionary perspective, in which evolved mechanisms like the 'flight-or-fight' response are inappropriately excited or suppressed, resulting in such conditions as anxiety or post traumatic stress disorders. Our own work suggests that sleep disorders may also be broadly developmental (Wallace, 2005b).

Serious mental disorders in humans are often comorbid among themselves – depression and anxiety, compulsive behaviors, psychotic ideation, etc. – and with serious chronic physical conditions such as coronary heart disease, atherosclerosis, diabetes, hypertension, dyslipidemia, and so on. These too are increasingly recognized as developmental in nature (see Wallace, 2004, 2005a for references), and are frequently compounded by behavioral problems like violence or substance use and abuse. Indeed, smoking, alcohol and drug addiction, compulsive eating, and the like, are often done as self-medication for the impacts of psychosocial and other stressors, constituting socially-induced 'risk behaviors' which synergistically accelerate a broad spectrum of mental and physical problems.

The picture, in humans, then, is of a multifactorial and broadly interpenetrating mind/body/social dysfunction, often having early onset and insidious, irregular, developmental progression. From the perspective of the Rate Distortion Theorem (Wallace, 2004; 2005a, Ch. 6), these disorders are, broadly speaking, distorted images of pathogenic external environments which are literally written upon the developing embryo, on the growing child, and on the maturing adult (Wallace, 2005a, Ch. 6). Equation (9) suggests that, in similar form, these images will be inevitably written upon the functioning of biological or machine consciousness as well.

Further consideration implies certain broad parallels with the development of cancer in multicellular organisms, a quintessential disorder of information transmission (Wallace et al., 2003).

The cancer model

Nunney (1999) suggests that in larger animals, whose lifespans are proportional to about the 4/10 power of their cell count, prevention of cancer in rapidly proliferating tissues becomes more difficult in proportion to their size. Cancer control requires the development of additional mechanisms and systems with increasing cell count to address tumorigenesis as body size increases – a synergistic effect of cell number and organism longevity.

As Nunney puts it,

> "This pattern may represent a real barrier to the evolution of large, long-lived animals and predicts that those that do evolve... have recruited additional controls [over those of smaller animals] to prevent cancer."

In particular different tissues may have evolved markedly different tumor control strategies. All of these, however, are likely to be energetically expensive, permeated with different complex signaling strategies, and subject to a multiplicity of reactions to signals.

Work by Thaler (1999) and Tellilion et al. (2001) suggests that the mutagenic effects associated with a cell sensing its environment and history could be as exquisitely regulated as transcription. Invocation of the Rate Distortion or Joint Asymptotic Equipartition Theorems in address of the mutator necessarily means that mutational variation comes to significantly reflect the grammar, syntax, and higher order structures of embedding environmental processes. This involves far more than a simple 'colored noise' – stochastic excursions about a deterministic 'spine' – and most certainly implies the need for exquisite regulation. Thus there are deep information theory arguments in favor of Thaler's speculation.

Thaler further argues that the immune system provides an example of a biological system which ignores conceptual boundaries between development and evolution.

Thaler specifically examines the meaning of the mutator for the biology of cancer, which, like the immune system it defies, is seen as involving both development and evolution.

Thus Thaler, in essence, looks at the effect of structured external stress on tumorigenesis and describes the 'local evolution' of cancer within a tissue in terms of a 'punctuated interpenetration' between a tumorigenic mutator mechanism and an embedding cognitive process of mutation control, including but transcending immune function.

The mutation control process constitutes the Darwinian selection pressure determining the fate of the (path dependent) output of a mutator mechanism. Externally-imposed and appropriately structured environmental signals then jointly increases mutation rate while decreasing mutation control effectiveness through an additional level of punctuated interpenetration. This is envisioned as a single, interlinked biological process.

Various authors have argued for 'non-reductionist' approaches to tumorigenesis (e.g. Baverstock (2000) and Waliszewski et al. (1998)), including psychosocial stressors as inherent to the process (Forlenza and Baum, 2000). What is clear is that, once a mutation has occurred, multiple systems must fail for tumorigenesis to proceed. It is well known that processes of DNA repair (e.g. Snow, 1997), programmed cell death – apoptosis – (e.g. Evans and Littlewood, 1998), and immune surveillance (e.g. Herberman, 1995) all act to redress cell mutation. The immune system is increasingly viewed as cognitive, and is known to be equipped with an array of possible remediations (Atlan and Chohen, 1998; Cohen, 2000). It is, then, possible to infer a larger, jointly-acting 'mutation control' process incorporating these and other cellular, systemic, and, in higer animals, social mechanisms. This clearly must involve comparison of developing cells with some internal model of what constitutes a 'normal' pattern, followed by a choice of response: none, repair, programmed cell death, or full-blown immune attack. The comparison with an internal picture of the world, with a subsequent choice from a response repertoire, is, as Atlan and Cohen (1998) point out, the essence of cognition.

One is led to propose, in the sense of equation (9), that a mutual information may be defined characterizing the interaction of a structured system of external selection pressures with the 'language' of cellular cognition effecting mutation control. Under the Joint Asymptotic Equipartition or Rate Distortion Theorems, that mutual information constitutes a splitting criterion for pairwise linked paths which may itself be punctuated and subject to sudden phase transitions.

Properly structured externally environmental signals can become jointly and synergistically linked both with cell mutation and with the cognitive process which attempts to redress cell mutation, enhancing the former, degrading the latter, and significantly raising the probability of successful tumorigenesis.

Raised rates of cellular mutation which quite literally reflect environmental pressure through selection's distorted mirror do not fit a cognitive paradigm: The adaptive mutator may propose, but selection disposes. However, the effect of structured environmental stress on both the mutator and on mutation control, which itself constitutes the selection pressure facing a clone of mutated cells, connects the mechanisms. Subsequent multiple evolutionary 'learning plateaus' (Wallace, 2002) representing the punctuated interpenetration between mutation control and clones of mutated cells constitute the stages of disease. Such stages arise in the context of an embedding system of environmental signals which, to use a Rate Distortion argument, literally writes an image of itself on all aspects of the disease.

These speculations are consistent with, but suggest extension of, a growing body of research. Kiecolt-Glaser et al. (2002), for example, discuss how chronic inflammation related to chronic stress has been linked with a spectrum of conditions associated with aging, including cardiovascular disease, osteoporosis, arthritis, type II diabetes, certain cancers, and other conditions. Dalgleish (1999, 2002) and others (O'Byrne and Dalgleish, 2001; Ridley, 1996) have argued at length that chronic immune activation and inflammation are closely related to the etiology of cancer and other diseases. As Balkwill and Mantovani (2001) put the matter, "If genetic damage is the 'match that lights the fire' of cancer, some types of inflammation may provide 'fuel that feeds the flames' ".

Dalgleish (1999) has suggested application of non-linear mathematics to examine the role of immune response in cancer etiology, viewing different phenotypic modes of the immune system – the Th1/Th2 dichotomy – as 'attractors' for chaotic processes related to tumorigenesis, and suggests therapeutic intervention to shift from Th2 to Th1. Such a shift in phenotype might well be viewed as a phase transition.

This analysis implies a complicated and subtle biology for cancer in higher animals, one in which external environmental 'messages' become convoluted with both pathogenic clone mutation and with an opposing, and possibly organ-specific, variety of tumor control strategies. In the face of such a biology, anti-inflammants (Coussens and Werb, 2002) and other 'magic bullet' interventions appear inadequate, a circumstance having implications for control of the aging of conscious machines which we infer from these examples.

Although chronic inflammation, related certainly to structured environmental stress, is likely to be a contributor to the enhancement of pathological mutation and the degradation of corrective response, it is unlikely to be the only such trigger. The constant cross-talk between central nervous, hormonal, immune, and tumor control systems in higher animals guarantees that the 'message' of the external environment will write itself upon the full realm of individual physiology in a highly plieotropic, punctuated, manner, with multifactorial impact on both cell clone mutation and tumor control.

Discussion and conclusions

These examples suggest that consciousness in higher animals, the quintessence of information processing, is accompanied by elaborate regulatory and corrective mechanisms, both internal and external. Some are already well known:

Sleep enables the consolidation and fixation in memory and semiautomatic mechanism of what has been consciously learned, and proper social interaction enhances mental fitness in humans. Other long-evolved processes probably act as correctives to keep Gilbert's evolutionary structures from going off the rails, e.g. attempting to limit flight-or-fight HPA responses to 'real' threats, and so on.

The implications of these examples for the control of progressive dysfunction in conscious or intelligent machines are not encouraging. 'Garbage collection' and other internal correction strategies are likely to be confronted by an interpenetrating and adaptable external system of 'virus writers' whose relentless selection pressure will continually challenge and degrade them. Irregular but progressive aging of conscious or 'autonomic' machines thus seems inevitable.

Animal consciousness has had the benefit of several hundred million years of evolution to develop the corrective and compensatory structures for its stability and efficiency over the life course. Researchers choosing to enter the next wave of artificial intelligence studies would do well to think very carefully indeed about the failure modes of conscious machines.

The explicit inference, then, is that machines which emulate human intelligence, and hence human consciousness, will likely suffer interpenetrating dysfunctions of mutual and reciprocal interaction with embedding environments which will have early onset and often insidious staged developmental progression, possibly according to a cancer model. There will be no reductionist 'bug in the program' whose 'fix' will correct the problem. On the contrary, the training of the machine, and its inevitable contact with the outside world, can be expected to initiate developmental disorders which will become more intrusive over time, most obviously following some damage accumulation model, but likely according to far more subtle, indeed punctuated, schemes. Retraining the machine would, of course, be possible, but might well obviate carefully learned behaviors whose supreme efficiency, after all, is the whole point of the device.

In sum, mission-critical machines designed to emulate biological systems are likely to fail insidiously, irregularly, and progressively, particularly when necessarily operating outside their training experience. These effects may well be quite subtle, manifest only at those rare but essential junctures where proper function is especially important, and indeed the basic rationale for the machine. In sum, the real world can be expected to write a distorted image of certain machine-specific pathogenic structures within it upon that machine, initiating a staged developmental path to critical failure.

Diagnosing and correcting such dysfunctions – 'machine psychiatry' – is likely to become a new and important engineering discipline.

This is not a new pattern. Radioactive waste disposal was considered a trivial afterthought in reactor design during the 1940's and 1950's, but, over just a few decades, became a chief bottleneck for the electric power industry. 'Program debugging' has grown to a principal impediment in conventional computer systems design, and the constant 'security flaws' and fixes of current operating systems are assuming legendary proportions. The second law of thermodynamics, it can be argued, dictates that, in spite of the Shannon Theorems, maintenance will be a principal bogey for all complex information enterprises – biological, social, or machine. Machine consciousness and 'autonomic computing' will be no exceptions.

References

Aiello W., F. Chung, and L. Lu, 2000, A random graph model for massive graphs, in *Proceedings of the 32nd Annual ACM Symposium on the Theory of Computing.*

Albert R., and A. Barabasi, 2002, Statistical mechanics of complex networks, *Reviews of Modern Physics*, 74:47-97.

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

Baars B., and S. Franklin, 2003, How conscious experience and working memory interact, *Trends in Cognitive Science*,

doi:10.1016/S1364-6613(03)00056-1.

Balkwill F. and Mantovani A. (2001) Inflammation and cancer: back to Virchow?, *Lancet*, 357:539-545.

Baverstock K. (2000), Radiation-induced genomic instability: a paradigm-breaking phenomenon and its relevance to environmentally induced cancer, *Mutation Research*, 454: 89-109.

Bennett M., and P. Hacker, 2003 *Philosophical Foundations of Neuroscience*, Blackwell Publishing, London.

Cohen I., 2000, *Tending Adam's Garden: Evolving the Cognitive Immune Self*, Academic Press, New York.

Corless R., G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, 1996, On the Lambert W function, *Advances in Computational Mathematics*, 4:329-359.

Coussens L. and Werb Z.(2002), Inflammation and Cancer, *Nature*, 420:860-867.

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

Dalgleish A. (1999), The relevance of non-linear mathematics (chaos theory) to the treatment of cancer, the role of the immune response and the potential for vaccines, *Quarterly Journal of Medicine*, 92:347-359.

Dalgleish A. and O'Byrne K. (2002), Chronic immune activation and inflammation in the pathogenesis of AIDS and cancer, *Advances in Cancer Research*, 84:231-276.

Dehaene S., and L. Naccache, 2001, Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework, *Cognition*, 79:1-37.

Dembo A., and O. Zeitouni, 1998, *Large Deviations: Techniques and Applications, 2nd Ed.*, Springer-Verlag, New York.

DSM-IV, 1994, *Diagnostic and Statistical Manual, fourth edition*, American Psychiatric Association.

Erdos P., and A. Renyi, 1960, On the evolution of random graphs, reprinted in *The Art of Counting*, 1973, 574-618 and in *Selected Papers of Alfred Renyi*, 1976, 482-525.

Evan G., and Littlewood T. (1998) A matter of life and cell death, *Science*, 281:1317-1322.

Forlenza M. and Baum A. (2000), Psychosocial influences on cancer progression: alternative cellular and molecular mechanisms, *Current Opinion in Psychiatry*, 13:639-645.

Gilbert P., 2001, Evolutionary approaches to psychopathology: the role of natural defenses, *Australian and New Zealand Journal of Psychiatry*, 35:17-27.

Grimmett G., and A. Stacey, 1998, Critical probabilities for site and bond percolation models, *The Annals of Probability*, 4:1788-1812.

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

Herberman R.(1995), Principles of tumor immunology in Murphy G., Lawrence W. and Lenhard R. (eds.), *American Cancer Society Textbook of Clinical Oncology*, ACS, Second Edition, pp. 1-9, 1995.

Khinchine A., 1957, *The Mathematical Foundations of Information Theory*, Dover Publications, New York.

Kiecolt-Glaser J., McGuier L., Robles T., and Glaser R. (2002), Emotions, morbidity, and mortality: new perspectives from psychoneuroimmunology, *Annual Review of Psychology*, 53:83-107.

Luczak T., 1990, *Random Structures and Algorithms*, 1:287.

Massimini M., F. Ferrarelli, R. Huber, S. Esser, H. Singh and G. Tononi, 2005, Neural activity spreads to distant areas of the brain in humans when awake but not when sleeping, *Science*, 309:2228-2232.

Molloy M., and B. Reed, 1995, A critical point for random graphs with a given degree sequence, *Random Structures and Algorithms*, 6:161-179.

Molloy M., and B. Reed, 1998, The size of the giant component of a random graph with a given degree sequence, *Combinatorics, Probability, and Computing*, 7:295-305.

Newman M., S. Strogatz, and D. Watts, 2001, Random graphs with arbitrary degree distributions and their applications, *Physical Review E*, 64:026118, 1-17.

Newell A,. 1990 *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA.

Newman M., 2003, Properties of highly clustered networks, arXiv:cond-mat/0303183v1.

Nunney L. (1999), Lineage selection and the evolution of multistage carcinogenesis, *Proceedings of the London Royal Society, B*, 266:493-498.

O'Byrne K. and Dalgleish A. (2001), Chronic immune activation and inflammation as the cause of malignancy, *British Journal of Cancer*, 85:473-483.

Ridley M. (1996), *Evolution*, Second Edition, Blackwell Science, Oxford, UK, 1996.

Savante J., Knuth, T. Luczak, and B. Pittel, 1993, The birth of the giant component, arXiv:math.PR/9310236v1.

Snow E. (1997), The role of DNA repair in development, *Reproductive Toxicology*, 11:353-365.

Somers S. and Guillou P. (1994), Tumor strategies for escaping immune control: implications for psychoimmunotherapy", in Lewis C., O'Sullivan C. and Barraclough J. (eds.), *The Psychoimmunology of Cancer: Mind and body in the fight for survival*, Oxford Medical Publishing, pp. 385-416.

Summit, 2005, 50th Anniversary Summit of Artificial Intelligence,

http://www.isi.imi.i.u-tokyo.ac.jp/ maxl/Events/ASAI50MV/index.html

Tenaillon O., Taddei F., Radman M. and Matic I. (2001) Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation, *Research in Microbiology*, 152:11-16.

Thaler D. (1999) Hereditary stability and variation in evolution and development, *Evolution and Development*, 1:113-122.

Volz, E., 2004, Random networks with tunable degree distribution and clustering, *Physical Review E*, 70:056115.

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. 2002, Adaptation, punctuation and information: a rate-distortion approach to non-cognitive 'learning plateaus' in evolutionary process, *Acta Biotheoretica*, 50:101-116.

Wallace R., D. Wallace and R.G. Wallace, 2003, Toward cultural oncology: the evolutionary information dynamics of cancer, *Open Systems and Information Dynamics*, 10:159-181.

Wallace R., 2004, Comorbidity and anticomorbidity: autocognitive developmental disorders of structured psychosocial stress, *Acta Biotheoretica*, 52:71-93.

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

Wallace R., 2005b, The sleep cycle: a mathematical analysis from a global workspace perspective, http://cogprints.org/4517/

Wallace R., 2005c, A modular network treatment of Baars' Global Workspace consciousness model, http://cogprints/4528/.

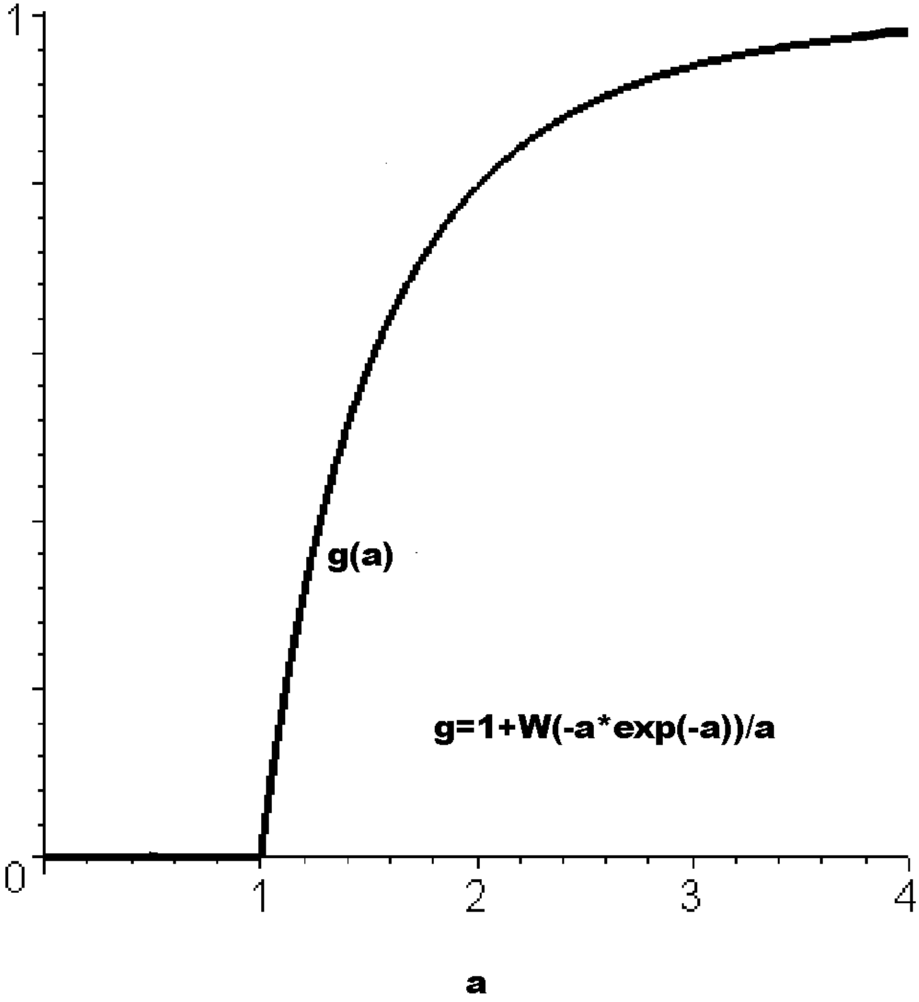
Waliszewski P., Molski M., and Konarski J. (1998) On the holistic approach in cellular and cancer biology: nonlinearity, complexity, and quasi-determinism of the dynamic cellular network, *Journal of Surgical Oncology*, 68:70-78.

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

Figure Caption

Figure 1. Relative size of the largest connected component of a random graph, as a function of 2× the average number of fixed-strength connections between vertices. W is the Lambert-W function, or the ProductLog in Mathematica, which solves the relation $W(x)\exp[W(x)] = x$. Note the sharp threshold at $a = 1$, and the subsequent topping-out.'Tuning' the giant component by changing topology generally leads to a family of similar curves, those having progressively higher threshold with correspondingly lower asymptotic limits (e.g. Newman et al., 2001, fig. 4).

RELATIVE SIZE OF LARGEST CONNECTED COMPONENT


g(a)
g=1+W(-a*exp(-a))/a
a