THE ROYAL SOCIETY

Levels and loops: the future of artificial intelligence and neuroscience

Anthony J. Bell

Interval Research Corporation, 1801 Page Mill Road, Palo Alto, CA 94304, USA

In discussing artificial intelligence and neuroscience, I will focus on two themes. The first is the universality of cycles (or loops): sets of variables that affect each other in such a way that any feed-forward account of causality and control, while informative, is misleading.

The second theme is based around the observation that a computer is an intrinsically dualistic entity, with its physical set-up designed so as not to interfere with its logical set-up, which executes the computation. The brain is different. When analysed empirically at several different levels (cellular, molecular), it appears that there is no satisfactory way to separate a physical brain model (or algorithm, or representation), from a physical implementational substrate. When program and implementation are inseparable and thus interfere with each other, a dualistic point-of-view is impossible. Forced by empiricism into a monistic perspective, the brain–mind appears as neither embodied by or embedded in physical reality, but rather as identical to physical reality.

This perspective has implications for the future of science and society. I will approach these from a negative point-of-view, by critiquing some of our millennial culture's popular projected futures.

Keywords: artificial intelligence; neuroscience; cyclic systems; dualism; science fiction

1. INTRODUCTION

In this paper I will survey the recent history, current status and future prospects of artificial intelligence (AI) and neuroscience. I will attempt to relate the social motivations and potential impact of the fields concerned on society at large.

2. THE SCIENCE FICTION FUTURE

Formalities over, and given that the Millennium is a significant enough social phenomenon that it colours popular impressions of the future of science, it is worth looking at what impressions a person of the year 2000 might have formed from late twentieth century popular science books, science fiction books and films, and even from the science pages of newspapers. Such a person might be forgiven for thinking that the future will be something like this.

> Nano-robots will perform all molecular repairs in our bodies, making us effectively immortal. Highly engineered drugs, perhaps the descendants of Prozac and Ecstasy, will take care of emotional disorders, as a side-effect solving all social problems, so everyone will be happy (finally).
>
> That's for the nostalgic minority who cling to living in the primitive biological form. More cyber-aware individuals will have downloaded themselves into the 'Net' and will exist like a William Gibson character in a global computer network which is capable of providing all protagonists with the most fantastic entertainment. Many global problems will be solved with the demographic move to the 'Net', problems such as population, food, transportation and energy.
>
> The 'Net-heads' will have been passed on the way by the 'Worldbots', digital mechanical life-forms which will first ease human life by performing all mundane tasks, but will shortly after become so much more intelligent than the unenhanced us that they will practically become 'spiritual machines', which may or may not use selfish altruism to decide to be benign towards the human animals, and if we are lucky, they will continue to serve us, something like digital Bodhisattvas.
>
> Back in the cyberworld, boundaries between individuals will break down, and transhuman life-forms will appear, analogously to the emergence of multicellular life in the ocean. Implanted into robot spaceships, these life-forms will lumber into space like the first amphibious fish lumbered onto the land. A long time after this, perhaps after a few galactic wars (in which the 'Dark Side' may be briefly flirted with but not joined forever), the universe will be one huge Internet, matter everywhere drawn into the process of computational living. The extremum of this is called the Omega Point. (A final twist is that since the Omega Point does not join the Dark Side, again possibly using game theoretic reasoning, it will decide to be benign and resurrect everyone who ever lived and give them what they most desire. This is called the Judeo-Christian heaven by Tipler (1995). Other references used in constructing this version of future history are Gibson (1986), Moravec (1990) and Kurzweil (1999).)

These amazing developments are the almost inevitable consequences of the merging of the digital and the organic worlds, on the threshold of which we are now standing. Cellphones and laptop computers are only the beginning. We might call this future the bio-informational age, in keeping with its millennial timing, and the smoothness with which it mixes in with elements of New Age philosophy.

3. THE CURRENT JOB OF SCIENCE

It's a giddy picture indeed, but how much of it, if any, will come true? If none of it is going to happen, it would be very helpful if science could tell us why, so that we could get on with living our real future.

The difficulty for science is that the prospect of a bio-informational future, with its cyborg, transpersonal themes causes us to ask questions concerning individuality, consciousness, mind and machine, exactly those questions which science has had least success in framing.

AI and neuroscience are the fields that come closest in engineering and biology to framing such questions. Scratch the surface of many AI researchers and neuroscientists (perhaps quite vigorously) and you may find someone who started off by asking 'What are we?'

The answers to this question are not that numerous. Either we are machines, in which case AI should be possible and neuroscience should be able to work out the algorithm (or algorithms) that the brain is running, or we are something else, in which case both projects will fail in their ultimate goals, which is not to say they will not achieve great things along the way. (One of the great things that they might achieve is an exact picture of their own limits.)

Either way, by examining the history and current state of AI and neuroscience and by identifying the issues beneath the surface of these fields, we may gather some sense of what are the important themes playing along science's internal frontier (disregarding for now how different this frontier looks from outside).

4. HISTORY AND STATE OF ARTIFICIAL INTELLIGENCE

AI's ultimate purpose is to build a robot that lives in the world with a computer for a brain. It therefore assumes that the essence of the living and/or thinking process can be captured in digital computation.

The first attempts to produce AI in the 1960s involved writing facts and rules into the machine using various quasi-logical languages. In the 1980s this became less popular. Rule-based systems were seen as non-robust: they could not adapt well to small changes in circumstances. Also, every fact had to be programmed in by a human. This led people to think that real-numbered, 'subsymbolic' systems were needed, and these systems had to be able to learn facts (or learn something) themselves, just by observing data. Historically, this view carried within it the cybernetics view of the 1950s.

It was one short step from this shift to statistical theories. The short step was called neural networks (Haykin 1999); it started in 1984 (Rummelhart & McClelland 1986) and it is not over yet. An interdisciplinary field with a higher than average tolerance for speculation and free-wheeling enquiry, neural networks were popular with students and military funders, and often regarded with frustration by other disciplines that shared a border. As the field became more rigorous, it re-established its connections with mainstream AI, through common interests in statistical machine learning. Technically speaking, the field of neural networks is contentless. The empirical side is neuroscience; the theoretical side is statistics and signal processing. This is perhaps what makes it such a great field to work in.

Symbolic AI was thus subverted by a shift to statistical learning theories. It was also subverted in two other directions by the emergence of the fields of artificial life (Langton 1997) and behaviour-based robotics (Arkin 1998) (or situated agents). Artificial life (or alife) is subsymbolic in that it implicitly assumes that intelligence is just the complex end of a simulatable life process. A living system and its environment are typically simulated together, often using genetic algorithms and population dynamics to simulate evolution.

Behaviour-based robotics attempts bravely to deal with the perceptual-motor loop of a robot in a real environment, rejecting both the alife simulated worlds and the mainstream AI notion of a representation of the world. Echoing Gibson (1979) in his famous debate with Marr (1982) (Bruce & Green 1990), the 'agents'-literature focuses on complex behaviour coming from simple mechanisms operating in tight coupling with a complex environment, in contrast to Marr's emphasis on the feed-forward computation of a representation from sensory data.

Alife and behaviour-based robotics lack a structural foundation such as that given to neural networks and statistical machine learning by mathematics. This makes it hard to judge progress or assess methodology in these fields. However, on the other side, neural networks that learn both sensory perceptions and motor actions in an environment are extremely rare, and for a good reason: it is difficult to build a statistical model of an environment when the system's perceptions are transformed into actions that affect the statistics of the input.

Furthermore, what should such an acting system do? There is an obvious goal for a feed-forward perceptual system: build a probability distribution of what happens. The hidden symmetries (dependencies, redundancies) in this distribution are the hidden structure of the world. But in this cyclic case, when the world is at least partly constructed by the actions of the system, the shape of this distribution is action dependent—the system gets to partly choose what symmetries exist, and the notion of a hidden set of privileged symmetries is under threat. This is post-modernism for statisticians.

At this point, most people would abandon informational, or unsupervised, goals and appeal to one of the many specific goals which a robot system might have, such as to find food or recharge the batteries. While these are no doubt important, they do have an air of arbitrariness about them that makes us uneasy: we are familiar enough with the flux of goals in our personal experiences to desire something more invariant to underly action selection.

5. QUESTIONS CURRENTLY LATENT IN ARTIFICIAL INTELLIGENCE

Here we have identified two questions which lie beneath the surface of the pluralistic AI of today.

The first question, to rephrase, asks why we do not have a mathematical theory of the perception–action cycle. Of course there is work on active perception, on sensory–motor coordinate systems, and engineering

department robotics is full of mathematics. But the kind of theory I mean is one that is as universally useful for characterizing cyclic systems as Shannon's information theory is for characterizing communications channels, i.e. feed-forward systems). (Incidentally, maximizing the channel capacity involves finding those hidden symmetries we mentioned that exist in the probability distribution of the input. This forms the basic goal of my own favoured area of neural networks—unsupervised learning (Hinton & Sejnowski 1999).)

Implicit in this is the second question. What would we want such a post-Shannon system to do? What quantity should a perception–action cycle system maximize, as a feed-forward channel might maximize its capacity?

A third question was directed at AI researchers by Penrose (1989), and by the hostility and controversy it caused, you knew he had hit a weak spot in AI. Penrose wondered if the fact that the physical substrate of the world, of which relativity and quantum mechanics are our best accounts, might be sufficiently different from the digital substrate of computers that it would render AI impossible. Is there something in the quantum that is necessary for mind?

Scoffing AI-philosophers characterized Penrose's position as 'we don't understand quantum mechanics and we don't understand consciousness, so they must be the same thing'. The derision increased when Penrose, to make his hypothesis more specific, proposed, with Stuart Hameroff, that quantum consciousness manifests itself through coherent quantum effects in a network of proteins called microtubules which form the structural skeleton of neurons (and other cells).

Critics, distracted by the strangeness of these specific proposals (which are not crucial to his argument), may miss the validity of Penrose's general doubt about the computer: that it is a particularly unusual artifact, being deterministic, discrete time and discrete state. The whole state of the machine at the digital level may be written down. No natural objects seem to be of this nature. The computer is really a physical instantiation of a model. We know a model can compute, but can it live or think?

Functionalism (the philosophy of AI) was based on using the computer metaphor for mind, arguing that the brain was the hardware implementation of the 'mental program'. But Penrose's arguments were really designed to raise doubts about this separation of physical and mental processes. Could the brain be separated from a supposedly finitely describable mental process running on it? Since René Descartes, the conceptual separation has been there in our language, but is it scientifically really there?

Either there is a physical level at which the separation can be performed (analogous to the level of logic gates in computers) or functionalists have to admit that the brain is not a machine. But the failure to detect a 'logic gate level' halfway up the brain's reductionist hierarchy may not be the end of the argument for the functionalist, who could still argue that if there is a computer at the bottom, AI would be possible, at the very least with a computer with the resources of the universe. The 'universe-as-computer' is a popular fringe-topic in physics, lying behind an effort to find a finite discrete process such as a cellular automaton that might underly the known laws of physics. But until someone succeeds in showing this, we might be wiser to stick with R. F. Feynman, who noted that quantum processes are not in general simulatable, even by Turing machines (and who in the process gave rise to the mysterious and unformed field known today as quantum computing).

The luck (or skill) of scientists is that sometimes they do not have to philosophize to find the answer. They can ask questions of Nature directly. So perhaps this is a good point to survey the history and current state of neuroscience, because this is the discipline whose empirical project is exactly the finite description of brain processes.

6. HISTORY AND STATE OF NEUROSCIENCE

The early landmarks in post-war neuroscience were the Nobel prize winning work of Hubel & Wiesel (1968) for their studies of the receptive fields of monkey visual cortical cells, and that of Hodgkin & Huxley (1952) for their uncovering of the mechanism and mathematics of spiking in neurons. It has grown into a huge field with the annual Society of Neurosciences meeting in the USA attracting 30 000 people.

The two early Nobel prizes reflect perhaps a natural split in the field between those working above or below the level of the cell. Many of the great successes of the 1970s and 1980s were at the subcellular level, as the molecular biology revolution progressed, and as a result this part of neurobiology was highly empirical and essentially continuous with mainstream cellular, molecular and developmental biology.

In this period, the molecular basis of neural signalling, both in spiking and synaptic transmission was uncovered. A bewildering array of ion channels, neurotransmitters and neuromodulators were found to be engaged in the processes of sculpting neural response properties and controlling communication between neurons. From the chemistry of photon absorption by photoreceptors, to the chemistry of muscle contraction, the nervous system apparently performed an astonishingly complicated and coordinated series of molecular actions not qualitatively different from those in other living cells, but somehow in the brain this molecular dance constituted percept, thought and action.

At and above the level of the spiking neuron, things were slightly different. Lacking the formal structural basis of molecular biology, neuron-level neuroscience focused on the spike trains as signals representing neural information. The discreteness of the spike as an information-carrying unit was matched in biology only by the genetic code. This led to early attempts to characterize the 'neural code', attempts that were revived by Bialek and co-workers in the 1990s (Rieke *et al.* 1997). (Notably, inevitably, these efforts attempt to characterize neurons as feed-forward information channels.) Behind these efforts is a faith in the neuron level, certainly as a useful descriptive level, but also as a 'computing level' which molecular and biophysical processes exist to implement. Does the goop that we see in the electron micrographs merely exist to implement 'the spiking computer'? This is the neuroscience analogue of the functionalist debate in AI, and I will return to it in § 7(c), after addressing the issue of cycles in neuroscience.

7. QUESTIONS CURRENTLY LATENT IN NEUROSCIENCE

(a) *Cycles in neuroscience*

The same problem with cycles presents itself in neuroscience as in AI, but whereas the primary cycle of concern in AI was the perception–action cycle, in neuroscience, the cycles are everywhere.

It is interesting that the clearest stories in neuroscience are those which at first glance most closely resemble feed-forward systems. One example is the synapse. The spike arrives at the presynaptic bouton, causing vesicles of neurotransmitter to be released, which in turn cause ion channels in the postsynaptic site to open and change the postsynaptic electrical potential. Another example is the early visual system, starting with the retina and moving through thalamus into early visual cortex. The treatment of this system as a feed-forward channel, despite massive corticothalamic and corticocortical feedback, has enabled information theoretic learning models the modest success of producing qualitatively correct predictions for the form of the static (Bell & Sejnowski 1997) and dynamic (Van Hateren & Van der Schaaf 1998) cortical receptive fields that were first observed by Hubel & Wiesel (1968).

However, feed-forward processing in the nervous system is the exception rather than the rule, and often what looks feed-forward contains complicated feedback systems at a different level of analysis. For example, the spikes of a cortical neuron have now been seen to extend far into the dendritic tree, affecting, through voltage-dependent channels, the integration of signals from synapses. This destroys the illusion that the neuron works like a directional 'neural network' neuron, performing a weighted sum of its input signals.

Even in the synapse and the retina there are feedbacks. Although the (human) retina receives no neural inputs from the brain, the brain controls gaze direction which determines what the retina sees. Although neurotransmitter does not travel backwards across synapses in most neurons, many other molecular signals do, as the extensive and controversial attempts to find synaptic Hebbian learning mechanisms in long-term potentiation have revealed.

In abstract, the lack of a theory of cycles in biology can be seen by considering an experiment in which some variable X is changed and some other variable Y is monitored. What is published are the relatively rare cases where some correlation in X and Y is observed. The temptation then is to say that 'X controls Y' and from this to build a model of feed-forward neural information processing (or if X is a chemical, we may market it as a drug to control Y).

In nature, things happen differently from in the experiment. X may rise, causing Y to rise, but then increased Y usually causes X to diminish, directly or through some other variables Z. These cycles of positive and negative feedback are universal in biology and cause equilibrium values of X and Y, or stereotypical dynamic behaviour to occur. A neural spike is one example of a transient dynamic caused by positive and negative feedback, where X is the sodium current and Y the potassium current.

Slipping into the language of probability theory, if we desire to discover the relationship in nature, of X and Y, we may measure their joint probability distribution $p(X, Y)$, and we could do so by observing X and Y under normal operating conditions, observing a peak in the distribution at equilibrium, and some trajectories corresponding to the stereotypical dynamics of the variables. But in trying to estimate whether X controls Y, experiments often take the form of measuring the conditional distribution $p(Y|X)$ and constructing the joint distribution through the formula $p(X, Y) = p(Y|X)p(X)$. This latter strategy gives the wrong answer for $p(X, Y)$ because (i) rather than the system controlling $p(X)$, we are controlling it, thus cutting the system at X, and (ii) we have, through our choice of independent and dependent variables, imposed on the system a direction $(X \rightarrow Y)$ of dependency, with an implied direction of causality that does not exist in nature.

There is no doubt that such experiments can still be useful in teasing out dynamic cyclic behaviour. The kinetics of ion channels can be identified with the aid of voltage and current clamping techniques, but there is a recognition in such experiments that the clamped cell is a frozen picture of the true process. This recognition often seems to go missing as the feedback loops get wider ('out of sight, out of mind') and particularly as biology becomes technology. Examples that spring to mind are the widespread prescription of drugs that combat depression by controlling seratonin levels, or attempts to control ecosystems by introducing new species, or, for that matter, the attempt to tailor many aspects of a plant's genetic make-up to fit an industrial model of agriculture. Anyone seriously studying or modelling metabolism or ecosystems knows the extent to which they are dealing with cycles, but somehow, when the results reach into the area of medicine or its macroscopic equivalent 'planet management', the causal, feed-forward style of thinking is what is presented, particularly to the news media and commercial interests. Anything which does not fit the feed-forward model is linguistically demoted to the status of a 'side-effect', to be eliminated if possible. But side-effects are nature's way of telling the scientist that all processes are cyclic.

(b) *Interlude: biology's master control node*

I cannot resist, at this point, discussing the role of biology's master control node, the genome. Although it is somewhat off the subject of AI and neuroscience, arguments pointing back to the genome as the causal factor behind animal behaviour and intelligence are so universal in our culture, that to allow the genome special status outside feedback cycles would be to endorse a control-node mysticism rivalled in shape and form only by that of the monotheistic Anglican bishops who debated so famously with T. H. Huxley. (When science became a greater authority on human origins than the church, the transition hid the fact that it was a change of government without a change in policy. Furthermore, affording the genome special status allows the present-day church of evolutionary psychology to rampage unchecked and, in my opinion, the wrong lessons are then drawn from biology.)

The genome's grand cycle with other genomes, mediated through populations of phenotypes is the king of all biological feedback loops. It is a trans-individual

molecular regulation loop, qualitatively similar to those occurring within cells, with cooperation (or symbiosis; Margulis & Sagan 1995) corresponding to the positive feedback loop and competition for resources corresponding to the negative feedback loop. Neo-Darwinists, stuck on the negative pole, like to interpret cooperative behaviour as 'selfish' altruism (I'll scratch your back if you scratch mine). The inverse position, on the positive pole, is to interpret competition for resources as selfless greediness (I'll eat you, but honestly, this is not about me). You might consider both positions absurd, or you might use the latter point of view as an antidote to the dominance of the former in our culture. The point here is that competition and cooperation have equal status and the process of 'natural selection' in which we are judged by an external environment (more biblical parallels) is better viewed as a complex molecular regulation loop like any other.

The regulation loop is mediated through phenotypic success, which brings up another loop-denying habit of neo-Darwinists, which is to see the genome as a controller for all aspects of the phenotype, right down to its specific behaviour: DNA as the determining code for an organism. There must be a particular attraction in this idea for certain authors, because they take great pleasure in outraging people's common sense by portraying organisms as the helpless puppets of their genes (Dawkins 1990).

I will not duplicate the effort of the many authors who have attacked the social or behavioural versions of this notion (for example, the preposterous notion that there could be a gene for homelessness, which was actually considered in an editorial in *Science*), because this would be to attack it at its weakest point. I'd like to attack the notion in its strongest version: the molecular. The central dogma of molecular biology is that 'genes make proteins, and not the other way round'.

The central dogma of molecular biology is wrong! Sequences of DNA code for strings of amino acids—true—but how these amino acids are assembled into functioning proteins and which parts of the DNA are read in the first place are both controlled by proteins, and depend on the state of the cell and its type. It's as if there was a bookish town (a cell) with a central library (the genome) and people (proteins) who came in to read short sections here and there, share with each other what they had read, and use the knowledge to build and change the town. Who is controlling here—the townsfolk or the library? (Answer: neither.)

Where did the people in the town come from? If 'genes make proteins', then the library made them, but the truth is that they were there all along. The functioning networks of enzymes that set to work on your DNA when you were conceived were already in place in the salty water of your mother's egg cell. They were just the latest instalment in a continuous epigenetic lineage that stretches back to your primordial metabolic ancestor, a droplet of seawater that accidentally got stuck inside a lipid membrane with a fortuitous set of amino acids.

It is harder to make more unsubstantiated assertions in biology than in the area known as 'origin of life'. But if the 'genes makes proteins' debate really comes down to whether there was RNA (code) before proteins (metabolism) or proteins before RNA in the first protocells (De Duve 1991), then two factors should be considered: (i) amino-acid chains form much more readily than nucleic-acid chains, and (ii) it is more likely that the first people wrote the first books, than that the first books wrote the first people. (It is noteworthy that both neo-Darwinists and New Testament theologians believe that 'in the beginning was the word (logos)'.) Of course, now it is claimed there were ribozymes (RNA with the ability to catalyse reactions), but was this metabolism evolving a code, or a code evolving metabolism?

The outcome of this debate is not crucial. The intent here is merely to weaken the notion of DNA as a kind of controller of the phenotype. An equally valid (and equally invalid) perspective has the phenotype choosing what is read from the gene and what is done with it. In reality, the organism and its genes are caught in a cyclic dynamic, and if the organism decides to spend its afternoon in a (real) library, instead of attempting to father children, then you can be sure that the pattern of gene expression will alter accordingly.

This argument fits with our first general theme of critiquing feed-forward thinking in AI and neuroscience.

(c) *Levels in neuroscience*

Returning now to the second theme we touched on when discussing AI, § 5 ended with a consideration of levels of a system and functionalism. There was a challenge to the functionalist to empirically investigate the brain and identify a level at which the brain could be finitely 'written down', a level analogous to logic gates in computers. The obvious candidate is the neuron level. If we wrote down the sequence of all spikes of all neurons, would that be enough to specify the 'neural computation'? Do molecular and biophysical processes exist to implement a 'spiking computer' at the neuron level?

I believe the answer to these questions is no. While no specific physical processes below the gate-level of a computer interfere with the model-like operation of the computer (unless something goes wrong), this cannot be said at the neuron level of the brain. Molecular and biophysical processes control the sensitivity of neurons to incoming spikes (both synaptic efficiency and post-synaptic responsivity), the excitability of the neuron to produce spikes, the patterns of spikes it can produce and the likelihood of new synapses forming (dynamic rewiring), to list only four of the most obvious interferences from the subneural level. Furthermore, trans-neural volume effects such as local electric fields and the transmembrane diffusion of nitric oxide have been seen to influence, respectively, coherent neural firing, and the delivery of energy (blood flow) to cells, the latter of which directly correlates with neural activity.

The list could go on. I believe that anyone who seriously studies neuromodulators, ion channels or synaptic mechanism and is honest, would have to reject the neuron level as a separate computing level, even while finding it to be a useful descriptive level. Perhaps a physicist or a neural-network theorist, in looking for an easy theory, would still argue that the molecular level is mere implementational detail, but in most cases this is more a result of prejudice, supported by laziness and ignorance. If the molecular level is unimportant for an

organism's behaviour, then how is a prokaryotic bacteria, vastly simpler than a neuron, able to navigate, eat and avoid toxins, all without the benefit of a nervous system?

If the neuron level is no good, are there any other candidate levels? Several have been proposed. The theory of neuronal groups, or cell assemblies, was another early candidate. The apparent 'noisiness' of individual spike trains could be smoothed out by integrating over groups of neurons coding, say, a given visual stimulus. The meaningful unit of perception was seen to be the activity of the group. In my view this idea contains a common error: failure to appreciate that noisiness is in the eye of the beholder, in this case the experimenter. In the case where a stimulus is presented and that part of the neural response which does not correlate with the stimulus is regarded as noise, we have a situation almost as bad as thinking French people are stupid because they produce strange noises in response to questioning.

What about the molecular level? Say we write down how many of each type of molecule are in each cell. Can this capture the computation of the cell? Unfortunately not, because the location of the molecules are important. Testing of enzyme reactions in bulk phase (solutions in test-tubes) is partly responsible for an impression that in the cell, molecules largely jitter around with Brownian motion and sometimes bump into each other and react. What turns out to be more likely is that most reactions take place locally in membrane-associated protein complexes, and the product of one reaction is passed directly on as substrate for the next. Evidence for this detailed spatial organization, called metabolic channelling, is accumulating (Ovadi 1995). Rather than being unreliable and 'wet', much of cellular biochemistry may already operate in what has been called the machine phase (although of course, in this paper I am arguing that 'machine', is the wrong word), where intricately detailed and coordinated reactions occur, not in the bulk phase. It seems that nanotechnology already exists, except that it is not technology in the normal sense in which a finite model is implemented using some particular substrate level. It is difficult to imagine human engineers making more efficient or complex processes by top-down manipulation of individual atoms.

We have reached the level of individual molecules, and the functionalist might say, no doubt through gritted teeth, that he is happy to write down the position of all the molecules in a brain. This will still be a finite description. If there is no evidence of submolecular interferences, we could have a 'molecular machine' to satisfy the functionalist. Remember that at this molecular level, we are looking for something as clean as a logic gate, which is a device responding deterministically to its logical inputs, and which is insensitive to the motions of individual electrons.

At this level, things become more controversial. Molecular computing is actually an area of advanced engineering research, so though it is not clear that it always falls within the discrete-state Turing model of computation, it might seem harder to dismiss the notion that molecules compute in nature.

If we use molecules to construct Turing-style computing devices, then, like good functionalists, we will have molecular computers. But what molecules do in nature may be different. In fact, it is. There are submolecular interferences that violate the separateness of the 'molecular machine' level, and they are quantum effects. Two examples of this are electron transfer in photosynthesis and the energetics of enzyme interactions (Welch 1986). In both cases, quantum coherences are necessary to explain the efficiency of the reactions.

But we don't even need to go as far down as quantum effects, because proteins do not end at the edges of the black and red balls of which ball-and-stick molecular models are constructed. Their electrical fields extend into the surrounding water molecules, orientating them to form what is called structured water. Structured water is also important in determining how enzyme reactions occur, and how ion channels are selective to certain ions.

To argue that one piece of structured water or one quantum coherence is a necessary detail in the functional description of the brain would clearly be ludicrous. But if, in every cell, molecules derive systematic functionality from these submolecular processes, if these processes are used all the time, all over the brain, to reflect, record and propagate spatio-temporal correlations of molecular fluctuations, to enhance or diminish the probabilities and specificities of reactions, then we have a situation qualitatively different from the logic gate. The variables lying beneath the level of a molecular 'gate' can affect the behaviour of the gate, so the functionalist is again frustrated, and the notion of the brain as a molecular 'computer' can be viewed as no more than an analogy, and an inaccurate one.

To say these things is not to be a 'New Age quantum mystic'. It is to attempt to clearly state empirical observations about molecular biology and to use them to attack the prevalent tendency to view biological organisms as machines in the exact technical sense in which computers are machines, i.e. in the sense that they are physical instantiations of finite models which do not permit physical interactions beneath the level of their machine parts (e.g. the logic gate) to influence their functionality.

It is a big leap from this argument to quantum consciousness. There is no evidence that large-scale macroscopic quantum coherences, such as those in superfluids and superconductors, occur in the brain. That some people like to make the quantum consciousness leap is testament more to the compelling connections between the mathematics of quantum mechanics and a holistic non-mechanistic world-view in which mind is immanent (Bohm 1980), than to any specific biological evidence. But as the first scientific workshops on 'quantum biology' meet, there is a good chance that a fascinating area of theoretical and experimental research will come about, and that more evidence will accumulate to suggest that functionalism cannot be used as a theory of the processes occuring in organisms.

8. RESTATEMENT OF THE ARGUMENT

In discussing AI and neuroscience, I have focused on two themes. The first is the universality of cycles, in other words of sets of variables that affect each other in such a way that any feed-forward account of causality and control is misleading.

The second theme is based around the observation that a computer is an intrinsically dualistic entity, with its physical set-up designed not to interfere with its logical set-up, which executes the computation. In empirical investigation, we find that the brain is not a dualistic entity. Computer and program may be two, but mind and brain are one. The brain is thus not a machine, meaning it is not a finite model (or computer) instantiated physically in such a way that the physical instantiation does not interfere with the execution of the model (or program).

9. THE BIO-INFORMATIONAL AGE REVISITED

What do these arguments say about the future, about science and society and their relationships? Will the cyber-dream take place, or should we quit AI and neuroscience and join a hippie commune? The technical conclusions on this seem to me to be as follows.

There will be no nanotechnological robots running around inside our bodies, at least none that are any more wizardly than the non-machine-like molecular complexes that already exist. There will be no 'control node' drugs that can pin us on the right end of the sadness–happiness spectrum, and thankfully we can drop this one-dimensional view of the human emotions. There will be no people living without brains, as digital patterns in the Internet. There will be no spiritual machines, models so advanced that they can deduce things that we find mysterious. There will be no machines with minds.

Cyborgs seem more plausible. The extension of human capacity through technology is already familiar to us, and it is a small step from driving a car to operating remote or tissue-embedded robot limbs. The process of building new models and surrounding ourselves with them will not be abolished in a return to some idealized pretechnological state that never existed. Models will merely be put in their place.

So if most of these things are not going to happen, where does society's focus on robots, virtual reality and the 'wired world' dream, come from? I believe it is a psychological reaction to the increasing proliferation of models around us. When social interactions become codified instead of open-ended, when people find themselves in roles as producers and consumers in a vast social machine, then the fantasy of the cyborg has already come true. When I enter an air-conditioned building in which the windows are all sealed and the lighting is all fluorescent, I am walking into a model, a virtual reality.

But the more our behaviour becomes machine-like, generated by and interpreted through the models that we and others construct, the more we will feel disconnected from the level below (and above) the models. We will be less able to see that we are not machines, and that there is no separating level at the logic gate that holds us above our physical substrate, and no control nodes in our brain that enable us to look down on reality. We are in the middle of it. I think this is a lesson that science is teaching us. If this lesson were truly to percolate into our culture from our science, and not be perceived by science as 'the threat of irrationality', then we would suddenly find ourselves living in a different world.

This is why I am ultimately optimistic about prospects for AI and neuroscience, despite my negative predictions about the success of their ultimate goals. I. Newton's mechanistic world-view took a blow with the arrival of quantum physics, but almost a century later, we still have physicists. Physics, it turns out, does not need to be tied to mechanism (in the strict sense we have used in this paper, quantum mechanics is non-mechanical), and neither does biology.

Computer science, mathematics, probability theory: these are more tied up with the building of finite models, but they too have an intriguing role to play, for along the border of the set of all models lurk paradox and inconsistency, the 'universal solvents' (to use D. C. Dennett's phrase in a situation where it applies) that dissolve models. This is very interesting territory, first explored by K. Gödel, who showed, remarkably, that there are true things that can be said within a consistent model which the model itself cannot prove. But interesting half-dissolved models can be built along the frontier, models that give paradox the respect it deserves. Quantum physics is one such model. After all, paradox is not just something to be obliterated at first sight, or ignored. Rather, it is an information structure which tells us exactly the shape and form of the failure of a model. (*Ex falso quodlibet* is what logicians say to express their observation that in Boolean logic, from 'true and not-true', anything is provable. But if this was the end of the story, then how could a Zen koan be useful, how could it be about anything? In fact there are a whole array of non-Boolean logics and paraconsistent logics. Some are even used in AI, reflecting the fact that when people are asked 'Do you like Bill Clinton?' many of them want to say 'I don't know' (underdetermined) and 'I love him and hate him at the same time' (overdetermined).)

Paradox informs us about the failure of a model in a qualitatively different way than Bayesian theory tells us that the observed and the estimated distribution of some variable are different. This suggests to me that there is something below probability theory, which, because the Cox–Jaynes formalism of Bayesian probability theory is founded on Boolean logic, may well be reachable by generalizing logical structures to incorporate answers other than yes and no.

These speculations, together with the empirical arguments I have made in the rest of this paper, suggest that there is a very exciting role for AI and neuroscience to play in the next century. As G.-C. Rota, a mathematician and an advocate of Husserl, Heidegger and Wittgenstein, wrote,

> Even in our days of constantly predicted revolutions, it is difficult not to be led to an optimistic conclusion. The new sciences of the computer and the brain will validate the philosophers' theories. But what is more important, they will achieve a goal that philosophy has been unable to attain. They will deal the death-stroke to the age-old prejudices that have beset the concept of mind.
>
> (Rota 1990, p. 107)

AI and neuroscience are exactly placed where the deaths of dualism and feed-forward thinking are scheduled to take place. If these disciplines choose to participate in this shift, rather than cling to concepts that are not empirically supported, then there will be many interesting PhD theses to write.

Finally, so far I have left out one question: Will there be a transhuman age? For this there is a strong biological precedent in the two major steps in biological evolution. The first, the incorporation into eukaryotic bacteria of prokaryotic symbiotes, and the second, the emergence of multicellular life-forms from colonies of eukaryotes.

Hegel had a word, sublation, for the harmonic incorporation of components into a whole without destruction of their individual nature, and we are all familiar with the good feeling that comes from playing in a team. However, those who followed up on G. W. F. Hegel's visions helped construct the nightmarish machine-like political state of mid-century fascism, so we are right to feel nervous about any superorganism with a hierarchical (i.e. feed-forward, controllable) structure. Thankfully, unlike twentieth century broadcast media, the Internet provides a good, non-hierarchical model for future information flow and social creativity. It is not risking too much to predict that it will continue to be a profound stimulus for social change.

Will this lead, ultimately, to some form of transhuman phase transition in the coming centuries? I believe that something like this may happen, and that science (and technology in some form, as with the Internet) will play a part in this. But I believe that at least part of this development will be a return to the past, a re-enchantment, to a vision of life that does not view humans or their minds as outside nature. Both our nostalgia for the past and our millennial fascination with a global cyber-reawakening are symptoms of the fact that we in the western world currently live in the most individualistic culture in human history. Our transhuman imagined science-fiction future may be, at base, a projection which contains the diagnosis of the present, as Jung might have observed.

Just like our private dreams, our public dreams are not to be taken literally. They are symbolic and indicative of imbalances in the present. The relieving news is that in correcting these imbalances, we will create a future which is not as as alien as the science-fiction future seems. In fact, it might look as familiar to us as something which we had forgotten.

REFERENCES

Arkin, R. C. 1998 *Behavior-based robotics* (*intelligent robots and autonomous agents*). Cambridge, MA: MIT Press.

Bell, A. J. & Sejnowski, T. J. 1997 The independent components of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338.

Bohm, D. 1980 *Wholeness and the implicate order*. London: Routledge and Kegan Paul.

Bruce, V. & Green, P. 1990 *Visual perception: physiology, psychology, and ecology*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dawkins, R. 1990 *The selfish gene*. Oxford University Press.

De Duve, C. 1991 *Blueprint for a cell*. London: Portland Press.

Gibson, J. J. 1979 *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Gibson, W. 1986 *Neuromancer*. Phantasia Press.

Haykin, S. S. 1999 *Neural networks: a comprehensive foundation*, 2nd edn. New Jersey: Prentice-Hall.

Hinton, G. E. & Sejnowski, T. J. 1999 *Unsupervised learning: foundations of neural computation*. Cambridge, MA: MIT Press.

Hodgkin, A. L. & Huxley, A. F. 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544.

Hubel, D. H. & Wiesel, T. N. 1968 Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–244.

Kurzweil, R. 1999 *The age of spiritual machines: when computers exceed human intelligence*. New York: Viking Press.

Langton, C. G. 1997 *Artificial life: an overview*. Cambridge, MA: Bradford Books, MIT Press.

Margulis, L. & Sagan, D. 1995 *What is life?* London: Weidenfeld and Nicolson.

Marr, D. 1982 *Vision*. New York: Freeman.

Moravec, H. 1990 *Mind children: the future of robot and human intelligence*. Cambridge, MA: Harvard University Press.

Ovadi, J. 1995 *Cell architecture and metabolic channeling*. Austin, TX: Landes; New York: Springer.

Penrose, R. 1989 *The emperor's new mind*. Oxford University Press.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. 1997 *Spikes: exploring the neural code*. Cambridge, MA: MIT Press.

Rota, G.-C. (ed.) 1997 Philosophy and computer science. In *Indiscrete thoughts*, pp. 104–107. Boston, MA: Birkhäuser.

Rumelhart, D. E. & McClelland, J. L. 1986 *Parallel distributed processing: exploration in the microstructure of cognition*. Cambridge, MA: MIT Press.

Tipler, F. J. 1995 *The physics of immortality: modern cosmology, God and the resurrection of the dead*. New York: Doubleday.

Van Hateren, J. H. & Van der Schaaf, A. 1998 Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond.* B **265**, 359–366.

Welch, G. R. (ed.) 1986 *The fluctuating enzyme. Nonequilibrium problems in the physical sciences and biology*, vol. 5. New York: Wiley.