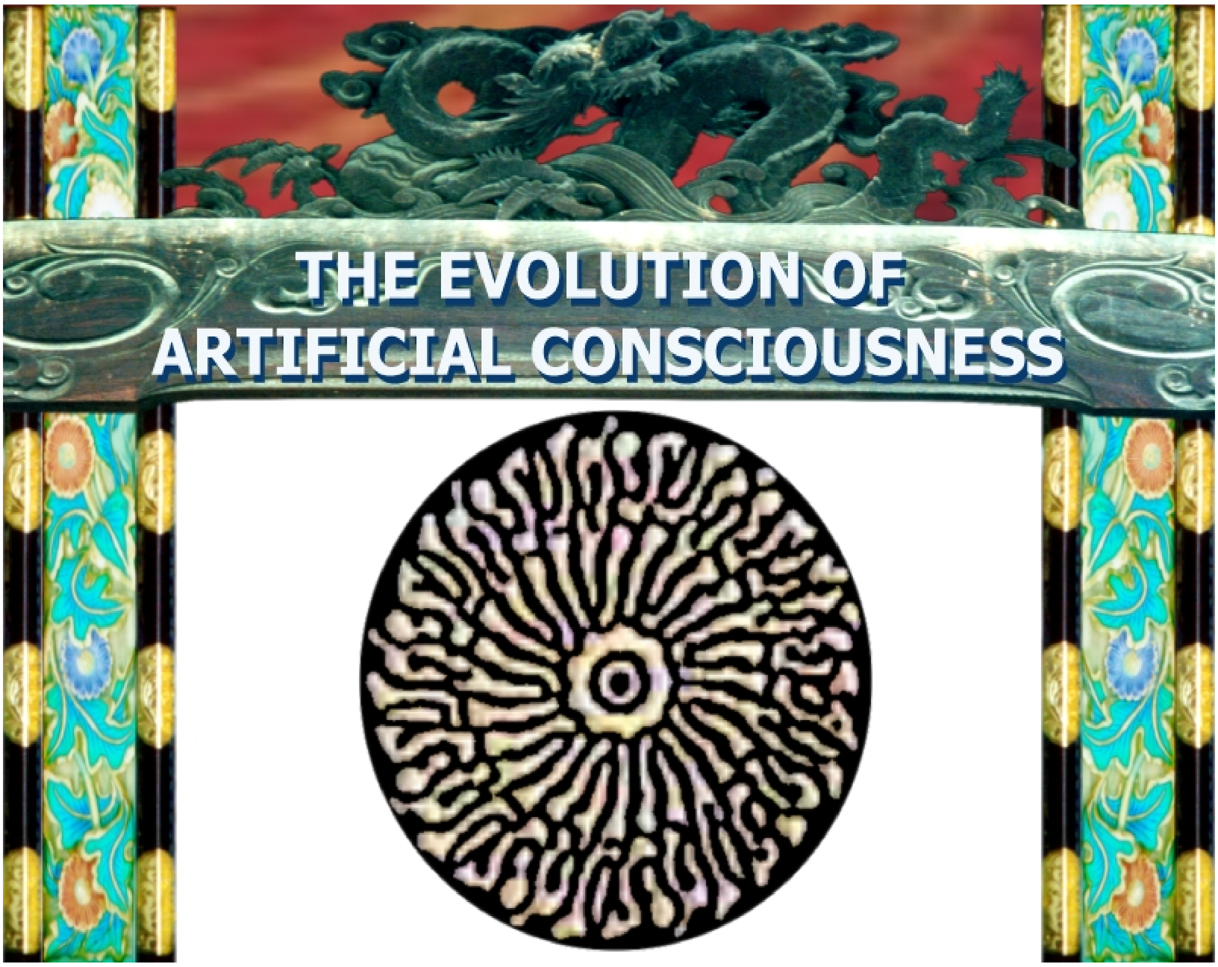


On the Evolution of Artificial Consciousness
Re-inventing the wheel
Re: Inventing the wheel

Stephen Jones
387 Riley St, Surry Hills
NSW 2010 Australia
sjones@culture.com.au

published in *Technoetic Arts*, vol.2, no.1, 2004

My intention here is to show how we might conceptualise the development of some kind of artificial consciousness.

The point about evolution that is worth recognising here is that it occurs in any system of interacting components over time in physical and chemical space, through sheer proximity, instability, variability, combination and recombination. These largely chemical processes can be brought on by variation in the energetics of the context (for example changes in photon flux affected by the diurnal cycle) in which the reactions/interactions take place and by variation in the proximity of "objects" (largely molecular). The kinds of interactions that these objects can undergo will produce resultant "objects" (molecules) which may then become useful in the catalysis of further reaction/interaction. Complexes of these now auto-catalytic objects (molecules) can then provide reaction/interaction components needed by other components extending the autocatalytic group. Somewhere along the line this autocatalytic process, as a self-organising dynamical process, leads to a system which can replicate itself. Thus life starts. Tom Ray defines evolution through this process:

> "Evolution is both a defining characteristic and the creative process of life itself. The living condition is a state that complex physical systems naturally flow into under certain conditions. It is a self-organising, self-perpetuating state of autocatalytically increasing complexity. The living component of the physical system quickly becomes the most complex part of the system, such that it reshapes the medium in its own image. Life then *evolves* adaptations predominantly in relation to the living components of the system, rather than the non-living components. Life *evolves* adaptations to itself." [Ray, 1994a, p181, my emphasis]

Evolution of organisms

Once life starts it proceeds by natural selection as described by Charles Darwin. Within any evolutionary system there are two overall kinds of components: organisms and the environment that houses the organisms, with which they interact in various ways. These two components have a complementary rather than oppositional relationship.

An organism is essentially a package of some sort, akin to the concept of a cell, containing a genome, which is a string of program codes (in DNA or some other instruction set, either preset or evolved) embedded in the package within some sort of support framework. This support framework provides the necessary components and structures to allow the string to be operated upon so that instructions can be executed. Genes are sets of instructions for how to build the organism and what to do with external resources and internal and external signals.

In evolutionary systems there will be some way(s) in which the instructions string can be modified either by random mutation or error, some form of cross-over (e.g. through sex) and in artificial systems by deliberate intervention from outside the system. Further it is necessary for the cell to be maintained and reproduced. [see Ray, 1994b]

Essentially an organism must operate within an environment. If the organism is able to adaptively find, recognise and use the stuff in its environment, then all else being equal, it stands a good chance of reproducing itself.

An environment is the container in which the genotype+package (ie. the cell) operates. This could be seen as the gene within the organism, or as the organism in its operating world space. It may carry other stuff such as "food" and (simulated) chemical products of the organisms' real or simulated metabolism. There will generally be a number of organisms of similar or different genotypes operating in the environment.

Environments are the space in which the genome is proved. This is what Natural Selection does. It is the process that determines whether the genome is successful at maintaining the organism and getting it to reproductive stage. In other words an environment is the space in which the organism is matured and uses its genes to survive by knowing or learning what to do with the resources and other things, including other organisms, contained in the environment.

Interaction

For any organism to have experience of, information about or knowledge of its environment it must run some sort of **sensory process** [Jones, 2000a]. Further, the primary way of having any effect on the environment is through the output of some kind of (by-)product which is a basic step in **interaction** of the organism with its environment. When other organisms in the environment respond to that output as though it were a probe then **communication** starts. Sensing, communication and the appearance of intentionality are the primary processes which all organisms (biological and simulated) will engage when they have any relation whatsoever to their environs. Mobility comes much later. I define sensing and communication, etc., as follows:

Sensing amounts to an "organism's" capacity to absorb difference relations from its context and to carry out such transforms (or processing) of those sense-data as to have them available as information about that context.

Sensors are required for sensing. They transduce whatever it is in the "world" that they are set up to handle. The result of this transduction is not so much a representation of the input as a transform from the input into the modes for which the sensor is enabled. This transform may then be presented to some sort of interpretive processor, and subsequently to an effector, or it may be directly "wired" to an effector or output device of some sort.

Effectors allow an organism to do things in the world, from getting out of the way (mobility) to catching prey to producing books or experimenting on the world. Effectors allow an output to be placed in the world-space which may be recognised as a signal by some other entity also within that world-space. This recognition will be species and experience dependent as determined by the kind of sensors available to that particular species of organism and the recognition filters which have developed through the individual organisms' experiential history.

Essentially these recognition **filters** are the meaning that the interpretive processor accords to the transforms it is presented. They will be embedded in the overall connectedness of the system as it has evolved through its experience, ie. the dynamics of its relations with the world (situatedness). Meaning can be partly a function of existential significance (say the recognition of food) or emotional significance or, say, intellectual/cultural significance. Recognition is also necessary for the establishment of communication.

Communication starts with the secretion or outputting of a probe or signal into the world-space in order to elicit a sensible response from other organisms (agents) in that context. When sensible and recognisable to another entity, which may or may not respond, a communication between entities may develop.

Intentionality may be said to appear when the sensory and/or communicative act is produced in the "direction" of an object in the context for the specific purpose (intention) of eliciting information from or about the object. Doing something, the use of an effector applied in the "direction" of the object of interest, may also be a useful sign of intentionality.

So organisms are productive, sensitive, responsive and communicative. It is these capacities as behaviour that afford the organism some form of intentionality and autonomy. They are the aspects of an organism's activity that will need to be developed in any artificial creature, AI or AC which we might want to develop. As such they form the basis for what is the most interesting result of evolution from our point of view, which is that it has produced nervous systems, brains and conscious beings which are capable of doing all sorts of things (not always particularly positive) in the environment. As John Taylor, of Kings College London, observes: "the evolution of multicellular organisms over the last 700 million years can be seen as the evolution of nerve nets in their appropriate "niche" environments." [Taylor, 1995]

Artificial Life

Artificial evolution is the operating procedure within artificial life. It is the experimental exploration of the processes of natural evolution through simulation, within computing systems, and as situated, hardware embodied, evolution of autonomous behaviour in robotic and similar systems. A variety of techniques, including genetic algorithms and cellular automata, are used to develop simulations and emulations of biological processes within computers, autonomous agents within the internet, robots and other artificial forms.

Essentially a genetic algorithm consists in a string of computer instructions in some robust instruction set within a purpose built computer operating system which functions as an ecology (usually fairly limited) providing the environment for the evolution of any artificial "creatures" which may be possible within the system. The paradigm form is the Tierra system invented by Tom Ray. [Ray 1994a, 1994b]

The evolution of a genetic algorithm produces more or less randomly modified versions of the original genome over a large number of generations with some process of natural selection editing out (killing) non-reproductive genotypes and mating successful ones. Mating produces new genotypes by mutation and sexual crossover of genes so that new genotypes are produced and attempt to survive in the environment of the time. The environment, of course, is usually some form of computer memory and is often pretty static, perhaps only providing food resources. Although there are now attempts to produce more dynamic environments having more ecological characteristics that place greater demands on the genotype for adaptive mutations. (This was the theme of the evolvability workshop at *ALife 7* this last August. [Nehaniv, 2000])

Each of the genotypes that appear will be a more or less successful solution to the problem of surviving within the environment. It is modification of the environment that draws solutions out of the algorithm. These solutions can be optimised for various kinds of technical computing problems, especially search algorithms, for behaviour development within specified or dynamic environments and for the development of "nervous systems" for producing suitable behaviour in robots in real environments.

Longterm product of robot and AI evolution.

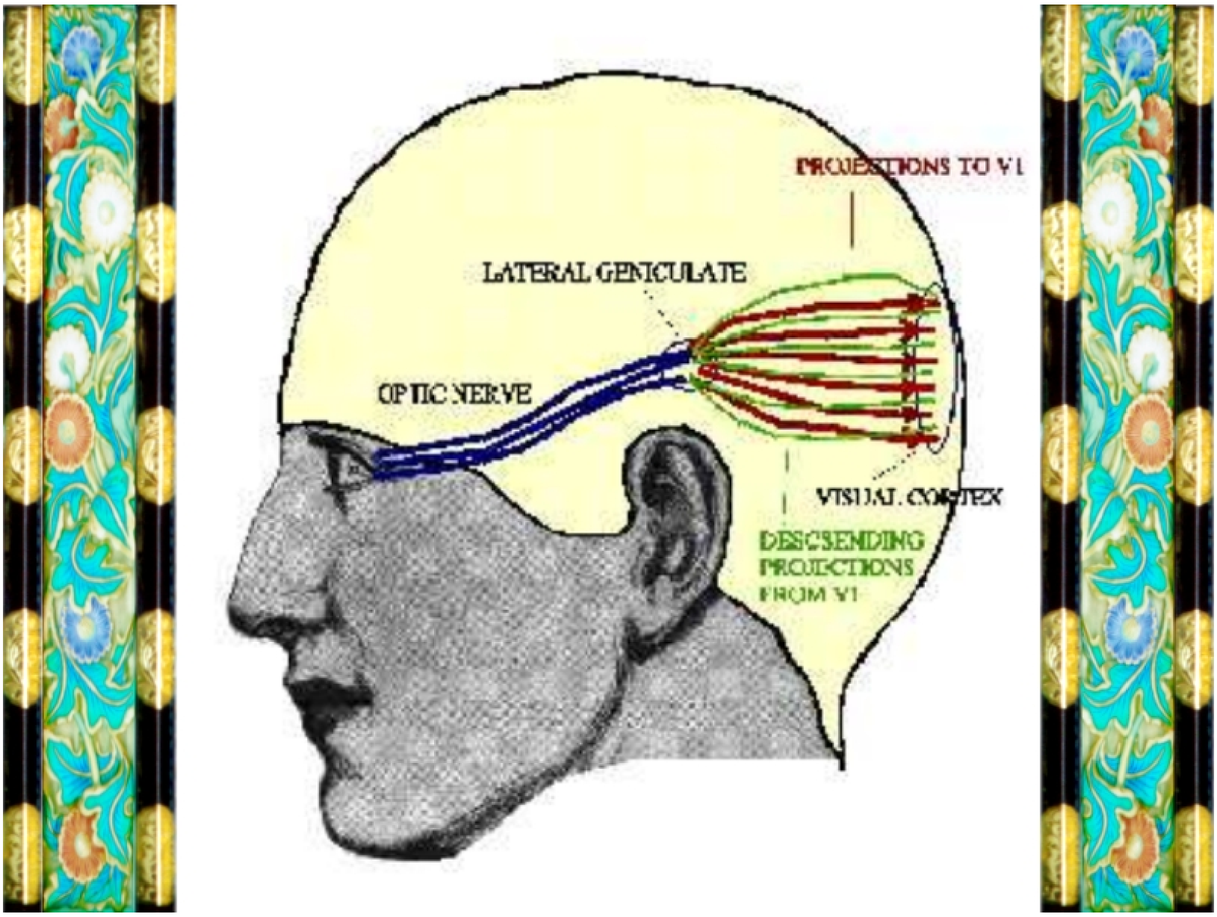
The use of genetic algorithms finds its most interesting use, as far as I am concerned, in the potential development of artificial intelligences and artificial consciousnesses. This sort of process starts in the evolution of robot control architectures, for example in work by Brooks [1999], Breazeal [1999] and others at MIT and by Husbands [1997] and Harvey [to appear] and others at Sussex University. Phil Husbands outlines the approach:

> "The artificial evolution approach maintains a population of viable genotypes (chromosomes), coding for control architectures. The genotypes are interbred according to a selection pressure, much as in standard genetic algorithm work. This is controlled by a task-oriented evaluation function: the better the robot performs its task the more evolutionarily favoured is its control architecture. Rather than attempting to hand design a system to perform a particular task or range of tasks well, the evolutionary approach allows their gradual emergence. There is no need for any assumptions about means to achieve a particular kind of behavior, as long as this behavior is directly or implicitly included in the evaluation function." [Husbands, 1997]

Given that it would take an excessively long period of synthetic evolution to develop anything useful through a "from-scratch" evolved genetic algorithm various short cuts have to be adopted. These involve at least the (pre)assembly, as "genes", of a set of instructions in some assembler-level computer language that do the work of, say, gathering data from the sensors, not unlike Ray's Tierra system [Ray, 1994a, 1994b]. Although we should recognise that Tierra is largely concerned with self-replicaton and only minimally engages with sensing. In this approach, a large part of the genotype is assembled according to some desired structure and then evolved to mould itself to the constraints of the particular task for which it is intended. In simple, carefully constrained environments, simulation evolved neural networks for handling sensors and effecting behaviour can be reasonably effective [Husbands, 1997]. But once the environments become more dynamic or the systems start to use complex sensing like vision, things become rather more complicated and the evolution has to take place in the actual physical robot in an actual environment so that the selection criteria are directly applied by the environment and all its messiness and dynamics are part of the criteria as applied.

Another approach might be to pre-establish the structures of sensors and effectors needed in these intelligent objects and then interconnect them as neural networks. The weightings on the connections, that is, the value placed on any connection, is then a function of its contribution to the successful behaviour of the object. Continued false alarms on an inappropriately connected warning system will tend to downgrade the usefulness of that connection while its appropriate connections to another kind of function will upgrade the value of that set of connections. Thus the organisation of the system will depend on use and experience within the operational context of the object. The development of weightings on connections need not depend on some oversight processor (usually a human). They could simply depend on some kind of foldback of negative responses inhibiting the connection versus positive responses reinforcing the connection as in the development and growth of the infant. This is the form known as a Hebbian learning network [Hebb, 1949]. With a mobile object an inappropriate response could be bumping into things, in an immobile object it could be opening an aperture (a window) too far thus flooding a photosensor. Husbands and Harvey again: "We advocate unrestricted recurrent real-time dynamical networks as one of the most general class[es] of behavior generating systems." [Husbands, 1997].

And it is at this point that I can start discussing the problem of evolving a consciousness or mind in an artificial organism, showing the kinds of behaviour I have previously discussed as indicating intentionality and affording autonomy in the individual entity. First we have to have some sort of idea about how to define consciousness. Like intelligence ultimately it has to be defined operationally. To paraphrase Brooks: consciousness (in some other individual) is in the mind of the beholder.



Characteristics of consciousness

We speak of ourselves as being conscious and we may allow that some other animals are conscious, though not to the degree that we are. We are also self-conscious and this produces a degree of difference from animal consciousnesses which may be what fully marks us out, that and our extraordinary language ability. Though there is increasing evidence that all apparent differences are more a matter of degree than of real novel capacities. [Language among parrots, tool using among jackdaws and among chimps, cultural traditions among macaques, see eg, Dautenhahn, 2000]

Within all conscious entities we tend to suppose intelligence, but intelligence doesn't necessarily make the agent conscious. All organisms behave autonomously displaying some level of intentionality and the capacity, by the use of sensors and effectors, to interact with their local environment (context). But Robert Kirk, of Nottingham University, [Kirk, 1996] suggests that directed or intentional interaction indicates consciousness in an organism when several conditions have been satisfied. First, the organism must be able to *use* the information. Next it must be able to *assess* the information for its usefulness which involves *interpreting* it. Then, it must be able to *initiate* and *control* its activity on the basis of the information it collects. And finally, particularly if the information is in some sense a signal, the organism must be able to *assess* the situation it is in so that it can decide how to *respond* or whether it should respond.

In us, our state of consciousness is always changing as we are exposed to continually novel sensations. We are selective of what we pay attention to. We posess a sense of unity across the senses and over time which the normal day-to-day changes of sleep and wakefulness, as well as abnormal changes such as unconsciousness, interrupt but do not render permanently unavailable. Paul Churchland, at the University of California at San Diego, [Churchland, 1996] adds several other aspects of consciousness, among them being short-term memory and its decay, conscious control over what we do, variable interpretation through thinking things through or reflection, the capacity to isolate sensory input in say, daydreaming and the disappearance of consciousness during sleep.

Moreover we possess an ability to respond to novelty in active and constructive ways. This is considered to be somehow over and above the mere 'irritability' of the senses and the body's reflexiveness to sensations which may or may not then have conscious impact. Further as our capacity with language demonstrates we have a generative capability by which we constantly produce new sentences and, analogously, new ideas and objects of cultural production.

So, to sum up our criteria. A conscious entity would need to:

Act independently and initiate procedures.

Do things such as gathering information for itself.

Do things which it initiates voluntarily for others.

Receive, store and reflect on perceptual input.

Show generativity, creativity and the capacity to construct.

Report upon the contents of our Consciousness

Interact with others in general, and

To project into the world for purposes of generating feedback.

But the main point that differentiates consciousness from that which we would consider to be entirely mechanical is our capacity for subjective or phenomenal experience. The experience of what it feels like to be something. It has been argued over the last 30 years (Nagel, 1974; McGinn, 1999; Chalmers, 1996) that we could quite happily do everything we do as conscious beings, that is display all the behaviour that we regularly do, without the subjective feel of the qualities (the "qualia") of our experience. This is the "Zombie" argument and it is used by philosophers such as David Chalmers to argue that consciousness must be something of a non-physical layer in our being, something that could not depend on our physical materiality for its presence, though obviously it needs the physical body to be in the world and acquire experiences for it to be subjective about. I, for one, am unable to subscribe to this view for a variety of reasons including the intrinsic relationship of experience to brain structure and the role of emotions and bodily chemistry, in the modulation of moods and feelings.

Now, I suppose that it is possible that we all do exist only as zombies without some sort of "soul" or extra capacity of a non-physical nature. One of the reasons this idea has developed is that it is extremely difficult for me to know that you have subjective experience despite the fact that I know I do. This is the basis of the privacy of consciousness, its incommunicability. I assume of course that you do have such subjectivity, but I cannot know for sure. But I do at least have your reporting of subjective experiences and your production, particularly of art objects, to suggest that it is perfectly reasonable for me to suppose that you do possess subjectivity.

But how could we possibly know whether a machine, no matter how good a reproduction of human behaviour it produces, could have anything like that subjective feel of being something that we possess? Perhaps we can only trust as we do with each other, in that the Turing Test really only requires that an AI be able, by its behaviour, its free interaction with us, to convince us that it is in fact conscious. For example the AI called "WinterMute" in William Gibson's *Neuromancer* trilogy ends up in the last book (*Mona Lisa Overdrive*) by removing itself from any interaction with the daily world so that it can produce Joseph Cornell boxes for its own pleasure. Something like this behaviour may be the only way we will ever be able to tell [Gibson, 1988].

The growth of consciousness in (human) organisms

So, by what process do we become conscious, or acquire consciousness? My feeling is that consciousness *accrues* as the infant matures. It is pretty clear that we don't come fully equipped at birth, but we certainly come with the capacity to become fully equipped over the early years of our lives. Of course some would suggest that we never become fully equipped or even fully mature, but...

The mind or the "I" is born empty of knowledge of the world. As John Locke [Locke, 1721] described it we are born "tabula rasa" (or a 'blank slate'). It is only by our experience of the world that we gain ideas of it. There are no "innate ideas". There may well be innate formal architectural and anatomical characteristics of a brain that determine just what subset of the possible aspects of the void it is that we are able to experience thereby contributing to the structure of our consciousness and to what we are enabled to perceive via our senses. But the anatomy needs experience to flesh out consciousness and provide meaning or interpretability to those things that it enables, thus making the world. [Jones, 2000c]

The evolution of the genome is not adequate to explain the development of the particular consciousness of any individual organism. It cannot specifically establish all the synaptic connections that the brain needs to function in the world. There are too many variables in the world. For example the fact that we have the capacity for language and the architecture to receive, understand and produce it is a function of the genome but the fact that we are able to acquire *any one* of the numerous languages human societies use, shows a massive flexibility of potential which could not be pre-specified. The genome is an algorithm for a body but algorithmic procedures cannot equip the body to handle the infinitude of variations in the world at any particular time. Adaptability and biological plasticity are essential.

The genome sets up the organism's anatomy and its physiological potentials but these have to be exercised by action and perception in the world (ie, by the production of experience) to come to any level of maturity, where maturity is thought of as being the capacity to differentiate self from other, to operate autonomously with intentionality and to understand whatever is necessary for the organism's survival in the world. In humans this has to be extended to our use of language and other sign systems, to our capacity to interact with knowledge and understanding and reportability in the social process of the culture. That is, the things we do that show us to be conscious are a function of in the world development.

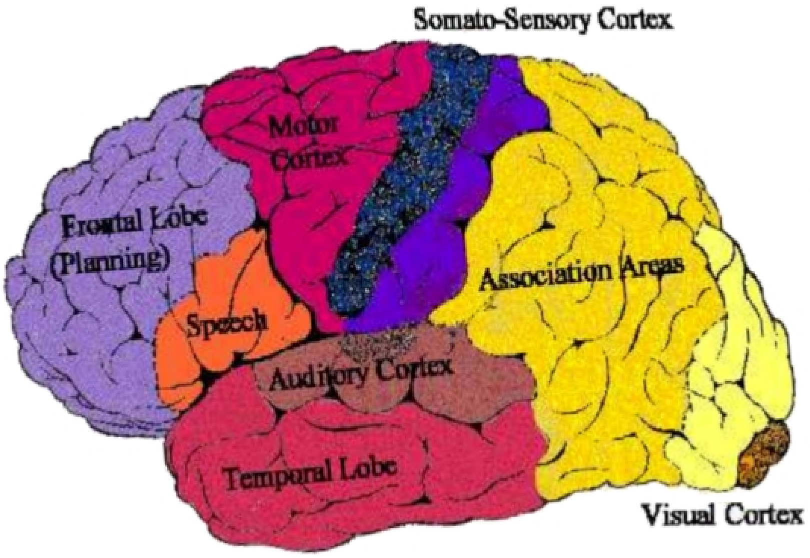
We *learn* to be in the world by our experience of the world as the brain and the rest of the development of our physiology is completed by the process of its exercise in the world, its contact with the world, the world's impact on it as a biological entity. Everything we know is a function of experience, either through sense perception or reflection upon that experience. It is the information produced from our senses that becomes the stuff of our subjectivity, that which we know.

This process of developing consciousness is epigenetic, that is it proceeds by stages [Zlatev, 2001]. Each level of the body and its knowing of its context has to be well established before the next one can commence. This is obvious to some extent, you can't learn about the world until the body is basically built through fetal development. Once born you can't see the world until the eyes are working, though this happens pretty quickly, within the first few minutes, but you can't understand what you see until you have the experience that provides the basis for that understanding. The muscles of the arms are triggered by various kinds of contacts with the world and these contacts in turn trigger the firing of neurons which set up synaptic transmission which connect neurons so that the next time the muscle can be used rather than just responded to. You can't run before you can walk, but to walk you have to able to stand, to stand you have to be able to pull yourself up, to do that you have to have control over your legs to do that it's really useful to do a lot of crawling and pushing against the floor and so on down.

Architecture of Consciousness

The brains of vertebrate organisms and especially conscious ones have a particular organisational structure which has evolved over a very long period, building on the structures of lower level organisms going back to the Cambrian. This organisation, or architecture, has been singularly successful in supporting the survival of particular groups of species leading ultimately to consciousness in humans and quite possibly to different kinds of "consciousness" in other large brained mammals such as dolphins and elephants, as well as the chimpanzee.

We can talk about this organisational structure, the architecture of consciousness, from two angles: (a) the anatomy and physiology of the brain and (b) the architecture of cognitive and associated functional activities. The physical body/brain is the biological substrate of consciousness, but its activities are sensing and experience, interpreting and reporting, calculating and emoting. These functional levels,



rational and emotional (not to mention the ecstatic and whatever other levels of consciousness we might have), that we call our *subjectivity* run on this anatomical architecture and to some extent (difficult to fully characterise) they have similar architectural relations to that which can be seen in the anatomy. These two regions of description make up the third person description - the brain - and the first person description - the mind - respectively.

The history of philosophy and the common sense of the last several centuries has promoted the divergence of these two descriptions, largely, one supposes, due to Descartes. But to all intents and purposes the mind needs the brain, ie, there is an anatomical substrate in which that subjectivity operates. The two layers of description can be mapped one onto the other and for our purposes here it makes sense to suggest that certain cognitive processes which we would want to implement in an artificial consciousness would need "bio-hardware" (or "wetware") that can carry out that process as part of an overall structure to which the cognitive process contributes.

A likely mapping between the cognitive architecture and the anatomical architecture makes its appearance through the work of Bernard Baars and the late James Newman. It seems clear that there is a reasonably direct and consistent mapping between these two layers of discourse, known as the global workspace [Baars, 1997a] and the thalamo-cortical loops system [Newman, 1997]. In fact it is what Newman and Baars had been doing in their general collaboration, allowing the development of the one model to guide the development of the other, in a bi-directional development process.

> Baars notes: "Cognitive architectures resemble theaters, typically taking input into a narrow "stage" of working memory, interacting with a large "audience" of semantic networks, automatic routines, and memory systems. This theoretical tradition has been systematically related to consciousness, at least qualitatively, in a framework called global workspace theory. For example, all cognitive architectures treat active elements in working memory as reportable; but reportability is the most widely used operational definition of conscious contents. Elements outside working memory are automatic or in longterm memory, and are therefore unreportable

> and unconscious. Cognitive architectures seem to reflect the same theater metaphor that is implicit in [Crick's] searchlight notion." [Baars, 1997b]

Baars' global workspace and his theatre metaphor of consciousness [Baars, 1997a] provide a model of the cognitive aspects for which there seems to be a growing consensus within neuroscience. While Newman's extensive reviews and synthesis of the neuro-biology of the thalamus and its ramifications into the cortex, coupled with the cortex's ramifications back into the thalamus and its various associated neuro-anatomical subsystems offer a model of the neural correlates of consciousness which is likewise achieving growing acceptance. Newman [Newman, 1997] has argued that this **thalamo-cortical system** linking to much of the cortex, plus the basal ganglia, the hypothalamus, the hippocampus and several other structures, functions as the system most likely to be the embodiment of the major processes of day-to-day consciousness and of the integration and control of the informational structure through which we have our place in the world.

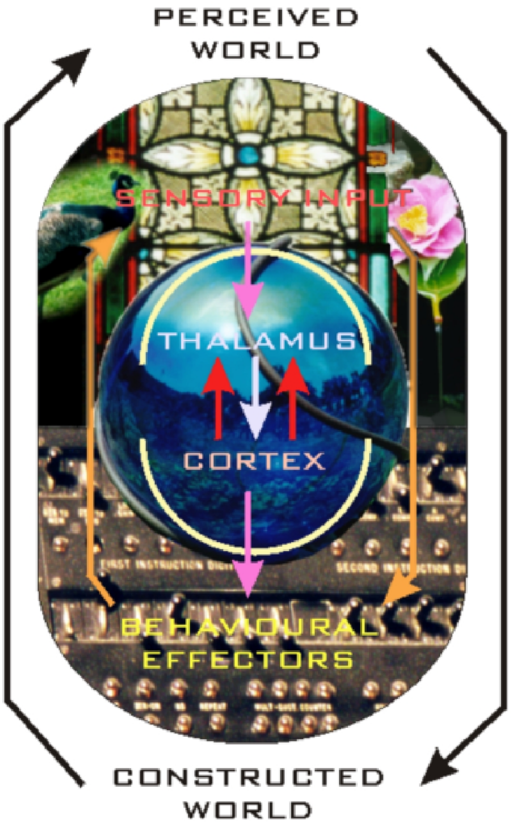
To summarise the thalamo-cortical system. The thalamus acts somewhat as the hub in a wheel, the spokes of which are nerve bundles traveling from the body periphery (carrying sensory data) and which are then relayed up into the cortex and cortical association areas for interpretive processing. [Newman, 1996] All of the sensory pathways (with the exception of the olfactory) are routed through the thalamus. For example, the optic tract runs from the retina, through the optic chiasm and thence into the lateral geniculate of the thalamus from where it is distributed into the occipital (or visual) cortex at the back of the brain. Auditory data from the inner ear is relayed through the medial geniculate into the auditory cortex in the temporal lobes. All of the face and body's proprioceptive data is routed through the thalamus on its way to the somato-sensory cortex. These are **ascending pathways**.

At the same time there are a vast array of nerve bundles descending from all areas of the cortex, particularly feature processing and pre-frontal control areas onto the intralaminar nuclei and the nucleus reticularis in the thalamus. These **descending pathways** act to gate the sensory data being presented to the cortex via inhibitory neurons in the nucleus reticularis. It is in this self-regulative capacity of the cortex to control what data it is being sent at any moment that we can find the function we call selective attention. In cognitive terms this is roughly the "searchlight" of Francis Crick [Crick, 1994] or Baars "spotlight" on the stage of the global workspace. There are also nerve bundles from the frontal and pre-frontal cortical areas going via the basal ganglia to the thalamus where they are integrated with sensory data to help in the control of motor functions.

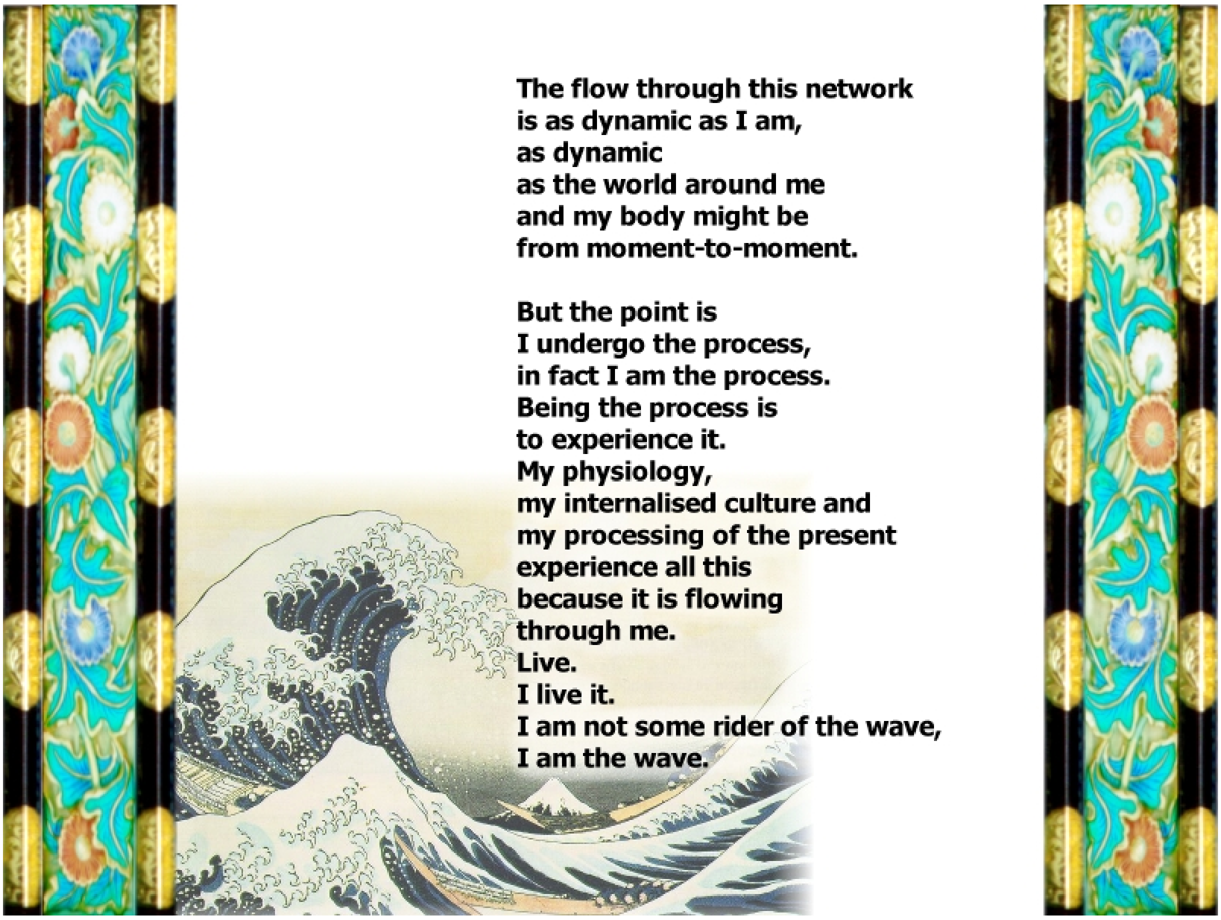
Illustratively, what's going on is that there is a massive array of reciprocally inter-connected **neural circuits**, **organised horizontally** around the thalamus and the basal ganglia and various emotion function nuclei giving behavioural control, and **hierarchically** between the cortex and the thalamus providing sensory control, especially useful in preventing the cortex from being overwhelmed by sensory input and effector control which can be driven from consciousness for learning how to do something and then relinquished to the lower level operational control once the practice has become automatic ("nailed down"), for example, in the difference between a beginner playing tennis versus a top seed player's capacity to return a 100kph serve with some accuracy.

Also, within the cortex there are vast arrays of horizontal inter-module nerve connections which pass feature processing transforms from one stage to the next as well as probably providing the capacity to associate different sensory modalities and to interpret grouped or bound collections of data from different senses that allow one to, for example, recognise that the sounds you hear are coming from the mouth you see speaking to you, and that the individual whose mouth you are watching is saying things that have meaning (or not, as the case may be).

One compelling piece of evidence suggesting the importance of the thalamus' role in consciousness. If the intralaminar nuclei of the thalamus suffer bilateral lesion damage then consciousness permanently ceases as happened in the case of Karen Quinlan.



It is in this thalamo-cortical loop system that a hardware basis for the working memory concept, Baars' Global Workspace might be seen, and it is this "hardware" architecture or some operationally similar version of it which will be needed in the development of an artificial consciousness. The thalamo-cortical loops system provides a means for integrating the flow of information in the brain, for providing selfregulated attention, for providing working memory through the memory that these feedback loops establish and for concentrating that information flow into the whirlpool of our conscious being. Subjectivity is our inside knowledge of what is currently being attended to (because it is our direct minded experience of that data flow) concentrated by the fact that it is that very centre of attention, what we are not engaged with is not an aspect of our subjectivity until it comes forthe and demands our attention.



**The flow through this network
is as dynamic as I am,
as dynamic
as the world around me
and my body might be
from moment-to-moment.**

**But the point is
I undergo the process,
in fact I am the process.
Being the process is
to experience it.
My physiology,
my internalised culture and
my processing of the present
experience all this
because it is flowing
through me.
Live.
I live it.
I am not some rider of the wave,
I am the wave.**

Implementation [Re: Inventing the Wheel]

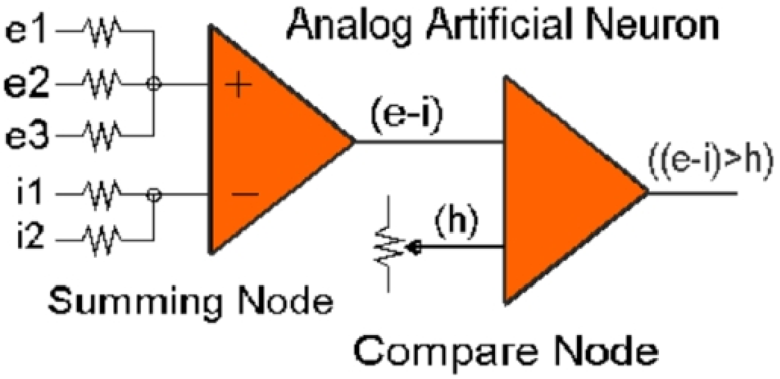
Obviously in a conscious brain a considerable amount of work goes on that is not conscious. It would of course be awfully inefficient if we actually had to pay attention to the smallest detail of each and every process in which we engage, not to mention deadly given that the heart might stop if we got distracted or our breathing might cease if we got too involved in, say, playing chess.

So a great deal of the processing of the brain goes on unattended by consciousness. This low level processing includes such things as getting the input from the retina into a state where it becomes seen and available for object recognition as part of the normal world, or getting what we hear into a state where it is interpreted it into language, let alone the process of making some sort of intelligent response to what is being said to us. But you may have noticed that when you are speaking what you are saying is almost completely unconscious, the stream of words goes through without deliberate intervention unless we suddenly notice that we said something wrong or we find that we have lost the word for some idea that we were trying to express or lost somebody's name. We listen to what we say and through this feedback consciousness cuts in, allowing us to correct ourselves or alter course as we go.

The cognitive is a hierarchically layered structure reflecting a hierarchical structure of the brain and implies, of course, that any machine consciousness we attempt to build must also be similarly organised. A great deal of the processing in the brain goes on unattended by consciousness.

So how do we implement this layered architecture? As in the development of the infant, hierarchically sequential establishment of processes of the body and its interaction with the world is necessary for setting up the correct ordering of mappings between behaviours and knowings. In robotic, or AC, development the first thing to do is to establish the low-level systems that allow the basic processes of sensing and maintenance to go on. We then need to produce the layers of feature detection and object recognition followed by control and planning systems. Each layer must be tied into the layers below with feedback systems which guide the activity of the lower processes so that they contribute to the needs of the system as a whole. Affective processing needs to be tied in here so that its motivations and the satisfaction of its needs are accommodated. Some sort of "hormonal process" or emulation of what is achieved in biological organisms through basic body chemistry hormonal control of its functioning would be one way of handling these factors.

But what technologies can we utilise to do this, to model the brain, both in organisation and in realisation? As already hinted at the brain is a multi-dimensional non-linear system of neural networks and thus it is neural networks to which we should turn. We will look at realisation first and then come to the organisation of what is realisable.

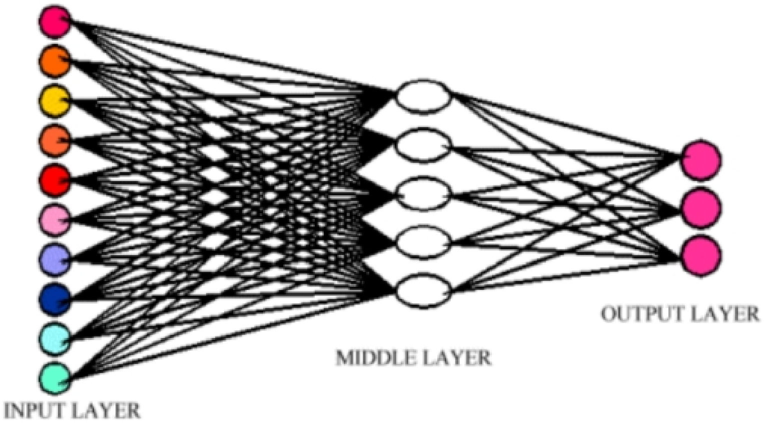


The first attempt to actually build some part of the human brain was McCulloch and Pitts' formulation of the artificial neuron in the early 1940's [McCulloch and Pitts, 1965]. They proposed a model of an artificial neuron which could be built electronically and would behave more or less as the biological neuron was thought to behave. It had a set of inputs - dendrites, a nucleus and an output - the axon. When the inputs received enough excitatory contacts they would sum to a voltage value above the firing threshold of the neuron and it would then release a pulse signal out through its axon.

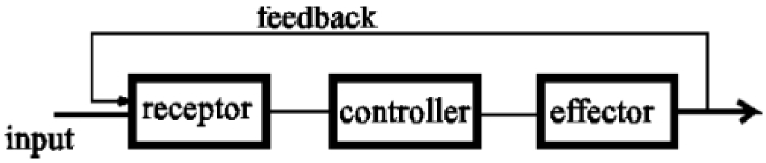
A neural network is a system of McCullogh and Pitts type artificial neurons, either simulated or in hardware, which are organised into layers. One layer is thought of as the input layer. This then feeds a layer or layers of processing neurons known as the hidden layers which in turn feed their results to the output layer. In most artificial systems there are only one or very few hidden or inner layers. It is very likely that the brain is organised as a very large structure of layers cascading one to the next as one progresses through various input, relay, subcortical and cortical neuron layers. These layers are generally very wide and in the cortical layers can have flexible numbers of neurons engaged at any one time in large assemblies which do the processing for features depending on how intense the input or the need to tease out detail in that input. Input can of course be from other layers within the system (reflexive) as much as it can be from specific sensory inputs.

Neural nets are connected up through synapses from one layer to the next. Mostly a neuron is thought of as having a large number of (dendritic) inputs and one output (via the axon) which can synapse on to a large number of input branches of other neurons in a series of cascaded layers. Each input synapse or connection has a value which is either positive (excitatory) or negative (inhibitory) and may range over some value say between 0 and 1. The value attached to a particular connection is called its connection weight.

Neural networks are able to learn and in fact must be set up by being trained to respond to input data sets (input stimuli) by presenting the data to the input layer and then reading the output and comparing it with what might be considered the appropriate output response to that input. The weights are then adjusted so that on the next presentation of the input the discrimination is a little finer or the result is a little closer to the desired result. This is the procedure in most neural nets used for engineering purposes, where the desired result is known. These are called classifier systems and are usually only feedforward networks where any adjustment of the weights is put in by hand, ie. by some external supervisory process. Some classifier systems do this by back propagation where the difference from the desired result is shown as an error value and this error value is summed back into the hidden layer connection weights so that they incrementally approach the values needed to fulfill the designer's intentions.



These feedforward neural nets build up a fixed look-up table (or dictionary) of input values to output values. But they tend to settle on a fixed solution and stay there and are not much like what happens in real brains and especially lack the necessary flexibility to handle real inputs in messy environments. Also in real brains the results cannot be known before hand so there is no way to build up a set of error values which can be fed back in. This then requires a different strategy for setting up the connection weights. Some sort of recurrence or feedback is required here. The transform of the input state produced by the network layer is not a grounded representation but the current state of the system. With further experience and the variability of the input data-set over its possible range we get a result which converges for any consistent series of inputs but shifts and develops as the input ranges across its possible values. In other words the result is drawn into the region of a solution, an attractor, but there is no *particular* fixed result of the input because, of course, in the real world no two inputs are ever identical and in a live system the current state is never ever as it was previously, simply through the accumulation of experiences. Feedback from the output, especially through other processing layers and back into the current input layer assists the network to converge onto a consistent result but provides a flexibility that allows us to say "well that looks like so and so, but it isn't quite them". Feedback is a kind of remembering, we produced a particular transform to this situation in previous experience and so we can reach for that transform again but with the added room to maneuver provided by other experiences in the meantime.



More usefully for conscious entities it is this kind of process that allows generativity. Kurt Godel's Incompleteness theorem [Godel, 1931] is often invoked as an argument against the possibility of producing a machine consciousness using symbolic processing procedures (GOFAI) [eg., see Penrose, 1989]. Godel showed that no formally specifiable system of procedures which is consistent can ever be complete. Because there is so much noise and ambiguity in the world, the specification, beforehand, of all possible ways of handling an input in all its possible states is doomed to failure. But Godel's result actually provides us with a hint towards the solution of this problem. We want a generative system that can handle ambiguity and works without knowing beforehand all possible results, a system which solves its problems on the fly. This is what the feedback layers do in the neural network system, they provide room for flexibility, for approximation, for noticing the similarity while at the same time identifying the pattern, the regularity.

In an open-ended, generative system, ie. a conscious system, the *actual* result of a neural net classification is not necessarily important, what is important is that the network generates more or less the same result for the same input and similar results for inputs of a similar type, so that there can be a spread of results, the changes in which match the changes in the inputs. This brings the whole problem of representation into a new view [see Hayles, 1991 or Harvey, 2000]. It is not important and quite unpredictable what exactly will represent, within the system, the state of any particular input. It is not what the representation is or matches that is important but that the possible transforms that the network does are consistent on repetition of the input and spread smoothly away from some particular transform as the input characteristics spread away from some notional paradigm version (the first impression) of the input, while possessing enough detail to make the necessary discriminations.

So my overall argument about what consciousness involves is that it is a function of myriads of looping connections in the brain. These feedback connections operate over a large range of layers of organisation. They may be within a sensory or effector modality, eg. all the connections within the visual system, or they may be between modalities so that, eg. speech and hearing stay in correct sequences and word formation matches vocal muscle control. Or they may be between a consciousness and its world and especially between individuals/others in the world as socially interacting beings, forming a wider set of feedbacks providing the basis for an entity's bootstrapping into the world, by bootstrapping the world into it.

Feedbacks loops allow information to flow from, say, the visual system to planning and control areas to motor effector areas and by the consequences of motor acts in the world back in through visual input and again to planning control to guide ourselves through the world. More tightly, feedback between vision and planning allow planning to regulate what aspects of vision it receives so that it can direct attention to something and get a better focus on some aspect of the scene. It is feedback loops in the structure of the brain that make it self-regulating. The brain regulates its own information flow-through so its various aspects get what they need for ongoing behaviour of the system.

Self-regulation is necessary for a biological system. To enable it to adapt to changing circumstances in its internal and external environment it must be able to learn and accommodate. Because there are no guaranteed preprogrammed solutions to changing problem spaces the system has to be able to develop its own solutions.

Every level of an adaptation is assisted by feedback loops. The system develops in an unpredictable context full of noise and the variability induced by other active agents in the environment. Any action the system makes will not be made in the same environment as any previous instances of that action, so it has to know what the effect of that action was so that it can vary it, preferably on the fly. This requires a short-loop feedback to know where the object which one was chasing has gone in the interval between starting the action and now. Feedbacks make an action adaptive to the new version of the context, especially since the context is changed by the action as well by other's actions. This is the observer effect. We perturb the world and then have to respond to our own perturbation. Longer loop feedbacks then tell us how others responded to the act, in an increasingly larger scale as we have impacts on larger and larger segments of society around us and on the culture. This is our memory of what happened helping us to make ongoing appropriate use of previously learned procedures in ever developing situations. This kind of layered feedback organisation is what makes consciousness a live process.

In neural network systems such as used by conscious beings, the training process during maturation is an intra-generational evolution. Aided by imitation of others - one's parents, teachers, older siblings etc. - by experiment, experience and the feedback provided by those others, one evolves adaptive behavioural solutions to the requirements of being in the world which are appropriate to the nature of your body in the real environment.

John Taylor again: "consciousness arises due to the active comparison of ongoing brain activity, stemming from external inputs in the various modalities, with somewhat similar past activity stored in semantic and episodic memory. The mental content of an experience therefore contains, as a component, the set of relations of that experience to stored memories of relevant past experiences." [Taylor, 1996]

Experience requires a body to have the experience and a context which carries what might be experienced. Embodiment is necessary for useful evolution so that functionally appropriate solutions can be arrived at in a dynamic living environment where nothing stays the same for very long. Bottom-up organisation grounds the entity in the world where its chemistry, cellular activity and the sensory and communicative processes of interactional behaviour make environmental resources useful and productions into that environment do-able and sensible. Feedback loops throughout provide the means, and autonomy, agency and intentionality come for free. When these requisites are sufficient and sufficiently organised consciousness itself is logically necessary and inevitable.

What then? [Re-inventing the Wheel]

Intentional, autonomous behaviour creates a bootstrapped relationship with the world. Bootstrapping in dynamical whole systems brings the condition of life. Living systems are a dynamic process, forming second order relations which I represent with the video feedback loop. It suggests how a simple physical system can show very elaborate emergence, much as a complex biological system can demonstrate something as elaborate as consciousness. Its dynamics imply time delays throughout the system and these make the system self-sustaining, traces of previous relations being present in the current "state" of the system as memory.

Ultimately, we are discussing the development of organisms and entities which would possess characteristics which we would consider as "human" in a bipedal primate. That is we are re-inventing the wheel.

What happens when we produce such an autonomous agent/robot with consciousness? When we provide it with sensors and effectors and the kinds of interconnectedness which provides it with the capacities to generate its own behaviour? The robot becomes an organism, a partner in the production of a cultural, ecological context, by which it itself can be affected just as we or any other organism will be. Essentially we are re-creating ourselves in other material forms. *Re-inventing the whee*l.

And, slightly adapting what Ted Kreuger has said: "Adaptive, intelligent and autonomous artifacts require a fundamental re-conceptualization of our relations with our [creatural] environments and one that indicates, as well, that we give some thought to the manner in which we are to work and communicate with them." [Kreuger, 1999]. Conscious entities of any sort must be considered to be high level organisms and must be accorded the ethical considerations that we ought to extend to all conscious organisms.

Any entities so constructed will not have anything particularly like human behaviour unless they are situated within a social context and embodied in a physical "body" which affords them the kind of interactions which human embodiment affords us. It is very likely that many kinds of conscious artificial organisms will not be embodied or situated at all like humans and will then have entirely different kinds of

formative experience and thus entirely different kinds of intelligence and behavior.

Whether we evolve or somehow directly construct these kinds of built organisms we must establish criteria for our behaviour towards them which acknowledge their autonomy and intelligence. They could well be considered as artificial intelligences or robots and ethical approaches like Asimov's Three Laws of Robotics, should be established as a very minimum. Also we need to consider how they will be trained or brought up, socialised and educated in being integrated into society. That is there may well need to be the kind of legal structures operating, like those in some societies these days, intended to prevent racism and similar kinds of discrimination. Given how little we have really achieved in these matters I dread to think what our attitude to intelligent or conscious artificial intelligences will be like in practice.

It is very clear from what I have said in the sections on developing such an entity that it will have to go through a considerable interval of training. The system will be constructed from some variety of neural networks (perhaps even heading towards real biological nervous systems) of very complex layering and interconnectedness. At initial start up the system will be effectively an infant and will need to be carefully nurtured and exposed to its environment of humans and other intelligent entities. It will need some kind of equivalent to parents and mentors. It will need to be "schooled" and socialised so that it can learn how to function effectively. Since these are situated, bottom-up constructions operating in the unpredictable real world there is little that can be done to *pre*-scribe all the things that it will come into contact with and have to be able to deal with. Such a system must be designed to accommodate to all sorts of eventualities, with the kind of flexibility that we normally associate with biological organisms. Perhaps in the long run the only way to make these entities is by using some sort of biological technology. Is this the real future of genetic manipulation?

Overall, the kind of approach that I am drawing out, means that the artificially conscious system will show responsive and communicative behaviour as well as intentionality and autonomy as an "organism". We must aim towards bringing the productions of their development processes into contact with human processes and activities so that the systems do not grow in isolation from the human environment. As I have argued elsewhere [Jones, 2000b]

> "to produce creatures which exist in some sort of isolation is, in the long run, to fail the creatures themselves and force them into a kind of confinement which we, under the same conditions, would consider cruel. In other words to become human requires the presence of other humans in the bootstrap into the cultural world and in the long run it would be bizarre to insist on complete isolation for any created entity. The kinds of entity that we might wish to keep at bay in this sense are produced under exactly those conditions of isolation in which we would become psychotic. Though the thought of this kind of future development might seem far fetched it is probably good prophylaxis to hold these matters in mind in the early stages of our attempts at artificial creation. If the enhancement and humanisation of communications and communication technologies is our goal then to do it under "de-humanised" conditions surely will breed failure."

References

Churchland, P.M. (1996) from his talk at *Towards a Science of Consciousness, 1996* summarised in The Brain Project, Jones, S. (ed) <http://www.culture.com.au/brain_proj/neur_net.htm>. See also Churchland, P.M. (1995) *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. MIT Press.

Baars, B.J. (1997a), *In the Theater of Consciousness: The Workspace of the Mind.* New York: Oxford University Press.

Baars, B.J. (1997b), "Metaphors of Consciousness and Attention in the Brain" *Trends in Neuroscience*,
Breazeal, C. (1999), "Robot in Society: Friend or Appliance?". In *Agents99* workshop on emotion-based agent architectures, Seattle, WA. 18-26.
Available at <http://www.ai.mit.edu/projects/sociable/publications.html>.

Brooks, R.A. (1999) *Cambrian Intelligence: The Early History of the new AI.* A Bradford Book, MIT Press, Cambridge, Mass.

Chalmers, D. (1996) *The Conscious Mind*, Oxford University Press.

Crick, F. (1994) *The Astonishing Hypothesis: The Scientific Search for the Soul.* New York: Charles Scribner's Sons.

Dautenhahn, K. (2000) "Evolvability, Culture and the Primate Social Brain", *Proceedings of the Evolvability Workshop at the Seventh International Conference on the Simulation and Synthesis of Living Systems* (Artificial Life VII), C. L. Nehaniv, editor, 1-2 August, 2000, pp. 23-26

Gibson, W., (1988) *Mona Lisa Overdrive,* Bantam Spectra,

Godel, K., (1931) "On Formally Undecideable Propositions of *Principia Mathematica* and Related SystemsI" in Davis, M.(ed) (1965) *The Undecidable*. Raven Press.

Hayles, N. K., (1991) "Constrained Constructivism: Locating Scientific Inquiry in the Theater of Representation" in *New Orleans Review, 18*, reprinted in *Realism and Representation: Essays on the Problem of Realism in Relation to Science, Literature, and Cultur*e (ed) George Levine.

Husbands, P., Harvey, I., Cliff, D. and Miller, G., (1997) "Artificial Evolution: A New Path for Artificial Intelligence?" *Brain and Cognition 34*, 130-159.

Harvey, I. (to appear) "Evolving Robot Consciousness: The Easy Problems and the Rest." to appear in *Evolving Consciousness*, G. Mulhauser (ed.), Advances in Consciousness Research Series, John Benjamins, Amsterdam. Available through <http://www.cogs.susx.ac.uk/users/inmanh/>

Harvey, I. (2000) "Robotics: Philosophy of Mind using a Screwdriver" in *Evolutionary Robotics: From Intelligent Robots to Artificial Life, Vol. III*, T. Gomi (ed), AAI Books, Ontario, Canada, pp. 207-230. Available through <http://www.cogs.susx.ac.uk/users/inmanh/>

Hebb, D. (1949) *The Organisation of Behavior.* Wiley.

Jones, S. (1998) "On Subjectivity" a prose poem from "Self Portraits from the Inside" in *The Brain Project* at <http://www.culture.com.au/brain_proj/new_pics.htm>

Jones, S. (2000a) "Sensing, Communication and Intentionality in Artificial Life." in Sugusaka, M. and Tanaka, H. (eds.) *Proc. 5th Int. Symposium on Artificial Life and Robotics*, Oita, Japan. pp26-29. Available from <sjones@culture.com.au>.

Jones, S. (2000b) "Intelligent Environments: Organisms or Objects?" to appear in *Convergence* spelcial issue on Intelligent Environments, ed Kreuger, T. Available from <sjones@culture.com.au>.

Jones, S. (2000c) "Bootstrapping the World into Being" in *Proceedings of Consciousness Reframed III.* Available from <sjones@culture.com.au>

Kirk, R., (1996) "On the Basic Package" Robert Kirk talks to Stephen Jones - The Brain Project <http://www.culture.com.au/brain_proj/kirk.htm>.

Kreuger, T. (1999) "Interfaces to Non-Symbolic Media." manuscript available at <http://www.comp.uark.edu/~tkreuger>

Locke, J. (1721) *An Essay on Human Understanding.*

McCulloch, W.S. and Pitts, W.H. "A Logical Calculus of the Ideas Immanent in Nervous Activity" in McCulloch, W.S. (1965) *Embodiments of Mind.* MIT Press.

McGinn, C. (1999) *The Mysterious Flame: Conscious Minds in a Material World.* New York, Basic Books

Nagel, T. (1974) "What Is it Like to Be a Bat" *Philosophical Review*, 1974, pp435-50.

Nehaniv, C.L. (2000) ed. *Proceedings of the Evolvability Workshop at the Seventh International Conference on the Simulation and Synthesis of Living Systems.* Technical Report No 351, Department of Computer Science, University of Hertfordshire.

Newman, J. (1997) "Putting the Puzzle Together: Part I: Toward a General Theory of the Neural Correlates of Consciousness." in *Journal of Consciousness Studies, 4*, No.1, 1997, pp47-66.

Newman, J., and Baars, B.J. (1996) "Bernie Baars and James Newman talk to Stephen Jones at Tucson II." available at <http://www.culture.com.au/brain_proj/baars.htm>.

Penrose, R. (1989) *The Emperor's New Mind.* Oxford University Press.

Ray, T.S. (1994a) "An Evolutionary Approach to Synthetic Biology: Zen and the Art of Creating Life." *Artificial Life 1*: 179-209

Ray, T.S. (1994b) "Evolution, complexity, entropy and artificial reality." *Physica D 75*: 239-263.

Taylor, J.G. (1995) "Towards the Ultimate Intelligent Machine", Presidential Address, World Congress on neural Networks, Washington DC, July 17-21.
Available through <http://www.mth.kcl.ac.uk/~jgtaylor/conscpub.htm>.

Taylor, J.G. (1996) "The Relational Mind", in *Neural Networks*, (ed) A Browne, Institute of Physics Press. Available through <http://www.mth.kcl.ac.uk/~jgtaylor/conscpub.htm>.

Zlatev, J. (2001) "Epigenesis of Meaning in Human Beings and possibly in Robots. to appear in *Minds and Machines*, 11.