

What must a brain do to be conscious?

Hugh Noble

Abstract

This paper brings together four ideas, three of which are well established: (i) subsumption architecture and its evolutionary implications, (ii) the idea that consciousness is associated with self-representation, (iii) the idea that the relationship between the mechanism of the brain and the subjective experience of consciousness is one of supervenience (not one of cause-and-effect). It also introduces a fourth which does not appear to have been suggested elsewhere, (iv) the idea that the evolutionary survival benefit associated with consciousness is that it enables an organism to anticipate its own reactions to predicted events. Based on these ideas, a speculative cognitive model is developed that would meet a minimum specification of what is required for a brain mechanism to be conscious. The model also offers an explanation of why we seem to view our own conscious experiences, from both inside and from outside our self-representation simultaneously and why the intuition of a mind-body duality is such a powerful and ubiquitous illusion. In addition, an approach to the processing of language is described, which is shown to be a plausible evolutionary development based on pre-existing features of the hypothetical brain-model.

Keywords: consciousness, supervenience, subsumption, model, self-representation, intuition

1. Introduction

It is disappointing that twenty five years after Marvin Minsky published "The Society of Mind" (Minsky 1985), human consciousness is still being discussed, by some, in terms which should have been consigned to a historical footnote - disappointing, but perhaps not altogether surprising. A physical explanation of consciousness has profound implications for popular belief in the nature of human identity as it also has for the architecture for AGI mechanisms.

The proposition being advanced here, endorses the view that the evolution of increasing intelligence enables a brain-mechanism to predict external events more accurately and to project those predictions further into the future.

A "theory of mind" (or ToM) augments that ability. It is the particular aspect of the mechanism which enables an organism to second-guess the likely behaviour of other animate beings.

What is being suggested here, however, is that in addition to that, consciousness is the particular aspect of intelligence that enables a brain-mechanism to anticipate its *OWN* likely response to predicted events. If that is the case, if being conscious is indeed associated with an identifiable survival benefit, then it is likely that it will occur in many other non-human mammalian species which have complex social lifestyles.

An argument which has often been raised against the idea of machine consciousness, is that human consciousness has the unusual and (according to those objectors), the *mysteriously inexplicable* characteristic of being "*subjective*". If the thesis advanced here is correct, however, that subjective quality has a simple explanation. Describing a personal

experience as "*subjective*" is merely an exercise in circularity. It is an alternative way of saying "*I am the mechanism which is performing that experience*".

To reinforce the idea of consciousness as a self-predicting mechanism, a hypothetical brain-model will be described here. The model does not purport to be a representation of the human brain in all its complexity, or even of a non-human biological brain. The model is a stripped-down, version of a thinking machine which can form concepts and can interpret its experiences in those terms. The model is given in functional terms only. It is an attempt to identify the basic minimum that a brain must be able to do in order to be conscious.

2. Preliminary Issues

Before describing the model some issues should be considered.

2.1 Assumptions

(1) EVOLUTION: It is assumed here that the mechanism of consciousness evolved, as did all other aspects of human anatomy and physiology, from earlier pre-conscious conditions which pre-date the advent of Mankind.

(2) SELF-REPRESENTATION: It is also assumed that a conscious brain has the ability to construct and utilise a mental representation of itself. We shall call this a Theory of SELFMIND (or ToSM) to correspond to a Theory of MIND (or ToM) which is concerned with an understanding and an ability to predict the behaviour of other animate entities. The need for a ToSM seems to be a well established principle and has been so for some time (McDermott 2001, McCarthy 1996, Minsky 1985).

(3) SUBSUMPTION: Rodney Brooks and others have demonstrated how a robotic device with subsumption architecture can function effectively without any centralized control (Brooks 1990). It is clear, therefore, that we do not need to address the otherwise puzzling question of how some form of strongly centralized control could have evolved.

(4) SUPERVENIENCE: It is argued here that the relationship between consciousness and the mechanism of a brain is one of supervenience. To speak of "*consciousness*" and of "*mechanism*" is to use two alternative ways of describing a single activity. When we speak about "mechanism" we are concerned with how the activity operates. When we speak about "consciousness" we are concerned about the kind of observable effects which the activity has on other systems and on the brain-mechanism itself. These terms are, however, actually referring to a single phenomenon, just as temperature of a physical body and the kinetic activity of its molecules, are really the same thing. They are not separate phenomena linked by some kind of causal relationship.

2.2 Comments on these assumptions

EVOLUTION: The notion, that human mental abilities have evolved slowly from earlier non-human brain characteristics, seems uncontroversial to anyone who is aware of the

biological, paleo-anthropological and bio-molecular evidence for evolution (and who understands that evidence). Nevertheless, there remains a stubborn school of thought that tries with determination to divorce consciousness from the implications of Darwinism (Penn et al 2008) - this despite a growing body of indirect evidence that non-human animals have some mental characteristics which are akin to consciousness. Chimps, for example, appear to have an ability to understand what other chimps are able to see and be aware of, and to anticipate how those other chimps will behave if they are aware of the location (say) of a tasty banana (Hare et al 2000). Macaque monkeys have been observed to be apparently willing to starve rather than be responsible for delivering an electric shock to their companions (Shapiro 2007). It is hard to explain these observations other than by assuming some form of conscious awareness of the situation and of the mental world of their companions. Objections to that conclusion are based on the idea that a Theory of Mind is a special, all-or-none, mental attribute which cannot be related to simpler mechanisms based on instinct and the observation of simple observable clues (Penn and Povinelli 2007). In this paper the contrary view will be advocated and supported by a model of how this evolutionary development could come about.

To demonstrate the feasibility of that idea, the hypothetical model of a conscious brain mechanism which will be presented here, will be developed a “pseudo-evolution” process.

PSEUDO-EVOLUTION: Pseudo-evolution is an attempt to reproduce the results of natural evolution while avoiding the profligate expenditure of resources, the enormous periods of time required and the unpredictable end-products of real evolution. The rules of pseudo-evolution are artificial but correspond to real evolution in significant ways. Development has to begin at a very simple level and then evolve in small steps. Each step must be the result of some small modification to an existing characteristic of the system, and each step must bring with it some immediate and identifiable survival benefit. The main difference between pseudo-evolution and the genuine version, is that pseudo-evolution is directed towards a particular destination, whereas natural evolution has no such intended directionality. It is true that occasionally the mutation of a single gene can modify the structure of the DNA molecule itself and in that way open up a greatly expanded landscape of possibilities. That can result in an apparent step increase in the rate of change. Evolutionary exploration of that new landscape, however, will still take place one step at a time.

EVOLUTIONARY DRIFT: Natural evolution can sometimes result in the development of features which have no recognisable survival benefit. This is often called "drift". It is a process of change which just happens. The only requirement is that the changes it produces are not actually detrimental to an organism's survival. What can be discounted, however, is the likelihood of drift producing a collection of features, which do not individually have any survival benefit, but which at some later stage suddenly find that in combination they offer some considerable and quite complicated advantage. That is simply implausible. It becomes increasingly improbable as the number of these co-operating components increases. To use drift as an explanation for the development of complex facilities will be termed here “miraculous drift”. And that is the main reason why the notion that "deep grammar” is an essential precursor to the development of linguistic communication, is here rejected. An alternative, more plausible development path will be proposed.

SELF-REPRESENTATION (A PROBLEM): Any process of self-representation involves the risk of infinite recursion. To avoid that computational impasse certain techniques must be adopted and these in turn have considerable implications which will play an important role in the proposed model of consciousness. We shall deal with these later (see Section 4.38)

SUBSUMPTION: Subsumption architecture has some important implications for how a brain mechanism might evolve. One consequence could be called "incremental augmentation". Instead of each mutational change *replacing* what went before, it simply *adds* a new facility to the system as a whole. The previous system continues to exist and to operate in parallel with the newly evolved parts of the system. This produces a situation in which old and new mechanisms are operating, to some extent, in competition, while at the same time they can rely on one another for specialist services. The brush strokes of evolution, we will argue, are visible in the landscape of the brain.

INCREMENTAL AUGMENTATION: "Incremental augmentation" should not be confused with other ideas, which have a similar terminology, and which are used in the field of evolutionary programming to describe various strategies which introduce gradual changes in the fitness criteria in order to lead the evolution of a system towards some desired solution to a particular practical problem (see for example Togelius 2004 or Koza 1990). That is a practical solution to the very slow rate of progress and unpredictable destination which are characteristic of genuine biological evolution.

SUPERVENIENCE: An understanding of supervenience is crucial to an understanding of the explanation of consciousness which is offered here. It is an issue which has been much discussed (Pro: Dennett 1990 and anti: Searl 1993, for example) and it is not as simple as it seems at first sight. The analogy with temperature and the kinetic activity of molecules is robust. But the situation is complicated by the fact that some writers use the term "supervenience" to cover a wider class of phenomena.

Supervenience and causation are often confused. The key difference is that of separability. The engine of a motor car *causes* that car to move forwards. But we can break that causal link by simply putting a foot on the clutch pedal, or by placing the drive wheels of the car in soft mud. But we cannot break the link between temperature and the kinetic activity of molecules. There is no link to break. If we modify one, we are, de facto, modifying the other.

Therefore, if the relationship between consciousness and brain procedure is indeed supervenience then it is pointless to ask how the physical brain mechanism could '*cause*' consciousness, or '*give rise*' to it, or '*generate*' it, or '*produce*' it. If we ask a question of that kind - or even if we ask what part of the brain activity '*correlates*' with consciousness (which leaves open the possibility of separation) - then our consideration of consciousness has gone wrong from the outset. It is also clear that the person who asks such a question has not been able to escape the grip of the Cartesian mind-body duality idea, no matter how vehemently they may deny it.

We can, of course, assign meanings to words in any way we choose and some have chosen a meaning for the term *supervenience*, which ignores or obscures the issue of separability. That is a pity because it throws away a useful distinction. Note too that if we construct a *simulation* of a supervenient relationship, that supervenient relationship can, and often will, disappear and be replaced by causally linked (simulated) conditions which

are generated by separate subroutines. In these (simulated) circumstances the two descriptions become two different procedures linked by subroutine calls. They can, therefore, be separated. So if we analyse a mechanism in terms of virtual machines, as does Sloman (1994), that essential property of inseparability will appear to have been lost.

In the context of this paper, the term “supervenience” is restricted to the inseparable form of relationship. To put the point, and its implications bluntly and in personal terms –

Consciousness is a procedure. I am an instance of that procedure. That procedure is ME. That procedure (which is ME,) by acting as it does, defines my identity, my individuality, my hopes and my fears.

When I take a decision and in so doing exercise my free will, that is ME (the procedure) taking a decision and performing an action which is based on my inherited predispositions, on the interplay between my inherited characteristics and my past experience, on my learned aptitudes, on my interpretation of events and on my memories of past events, particularly those which are currently easily accessible due to the accident of recent events and on my interpretation of those events.

When that procedure does whatever it does, I am the one who is doing that. I am the one who is doing (not having) an experience.

Those who are inclined to dismiss that idea as “mere clockwork determinism” should consider the monstrous complexity of the system, the fact that its behaviour is “chaotic” in the mathematical sense, the way it is subject to the impact of unpredictable external events, the way the nature of those events is subject to interpretation, and the way every reaction between the chemical molecules involved in brain activity is ultimately dependent upon unpredictable events at a quantum level. They could then try to predict what time will be showing on the face of a clock when the spring or battery runs down. Perhaps they should then reflect upon whether their view of “mere clockwork determinism” might be based on an overly simplistic notion of the nature of causation and a naïve concept of determinism.

AWARENESS: We are often reminded that awareness is not the same thing as consciousness. It is harder, however, to find, in the literature, any explicit explanation of what exactly is the difference between them. If we are to understand consciousness, however, we do need to understand awareness, how it operates, where it comes from, and at what level awareness enters the account of how consciousness evolved. The approach taken here is quite explicit.

Even at a very primitive level, living organisms are able to receive sensory input, identify particular patterns of signals, and then respond in a way that has been tested by natural selection and found to have survival benefit. Responses of that kind appear to be purposeful. However, no self-knowledge and no “awareness of being aware” is involved in this early form of behaviour. We might, nevertheless, call it “instantaneous awareness”. The word “instantaneous” refers to the duration of the phenomenon. It is gone almost before it exists. A sunflower could be said to have this degree of awareness, as it turns its head to face the sun. Even a thermostat could be said to have “instantaneous awareness” of ambient temperature. Instantaneous awareness is not a big deal.

But if that prosaic meaning of "awareness" is adopted, more significant forms of awareness can then evolve from it - slowly and in small steps. The important factor is what the mechanism does with the information it receives (and is instantaneously aware of).

That way of looking at awareness has great advantages. At some later point, when the system is more fully developed, we will not be faced by the need to introduce an apparently sudden step change in the system as (in some mysterious way), the system acquires the ability to be aware of what it is doing. Being aware, like being conscious, is a performance. It is the detail and elaboration of that performance which defines the degree of awareness.

3. The Pseudo-evolution of the Brain Model

Before describing the model in some detail we provide an outline "road-map" of the development process.

3.1 Phase-1: The Stimulus-Response Automaton

The first phase of this development provides us with a stimulus-response automaton (or SRA) which is capable of survival on its own. Its development starts with a rudimentary system, equivalent to a single-celled organism, which is not capable of doing anything more complicated than responding to changes in the strength of incident sunlight. To that simple system, evolution gradually adds, by incremental augmentation, more types of response to an ever widening range of potential stimuli, to become as complicated as (say) an insect. As it does this, it develops a number of additional features which have both an immediate use, and survival advantage, but which, at a later stage, also provide extra advantages.

3.2 Phase-2. Transitory and Persistent Memories

The second phase introduces a transitory memory. This enables the system to develop more complex responses to various stimuli. It could, for example, enable an organism to run or hide from a predator and continue to do so for some time after the predator is no longer visible. In this phase we also see the introduction of a second and more selective medium-term or persistent memory. This preserves important experiences and, significantly for further developments, provides the mechanism for selection and the repetition of patterns of behaviour which achieve beneficial goals.

3.3 Phase-3: Concept Formation

The third phase introduces the processing of the persistent memory in order to compress it. As a side effect, this produces a collection of chunks of sensory data which occur

repeatedly within the memory store. This is the first stage in the development of "concepts". Within phase-3 the system is then able to utilise these concepts to interpret on-going sensory experience. Dream-sleep is also introduced at this point as the need to process memory becomes more demanding and cannot easily be carried out while the need to process current input is on-going. The compression routine is then re-applied to concepts, which have already been formed. This creates abstract concepts.

3.4 Phase-4: Mind and SelfMind.

At the fourth stage of development, the concepts of MIND and SELFMIND are introduced. These are products of the additional application of compression process, to overt clues to predictable animate behaviour (body language, facial expressions, gestures, tone of voice, eye-pointing). Having formed these concepts of MIND and SELFMIND, the system is able to use them as a means to distinguish between (i) what is believed (and was believed to be known in the past) (ii) what is currently "known" (or believed to be true) and (iii) what is believed that other people believe. It is at this fourth stage, when the system is processing data at a more sophisticated conceptual level, storing the interpreted stream of experience within the SELFMIND, and is able to relate these experiences to its own goals and anti-goals (made available to it by the phase-1 SRA) that we can legitimately talk about the system being "conscious".

3.5 Phase-5: Language

The fifth stage takes us beyond a basic level of consciousness to a higher level of consciousness made possible by the use of a spoken language to facilitate intercommunication, to provide labels for abstract terms, and to bring the concepts possessed by various individuals, into a greater degree of correspondence.

4. The Brain Model

In its earlier stages, the development of this hypothetical model could be described as unremarkable because it does little more than follow a path well known to AI researchers. It is only in the later stages that some novel aspects appear. Nevertheless the system as a whole is crucially dependent upon those early stages. It is necessary therefore, to go through the development in its proper sequence to show that the development is natural and entirely compatible with (real) evolution.

4.1 PHASE-1: The Stimulus-Response Automaton.

The first stage in the development of the proposed system can be described as a stimulus-response automaton (or SRA). We can envisage a sensory array which is the source of

input signals, and a response array which is the destination of the SRA's output signals. Between these are a number of perception-units organized in a network of intercommunications. Each perception-unit has the task of recognising a particular pattern of signals and responding by generating one or more output signals.

Since our aim here is to demonstrate only the feasibility of artificial consciousness and the model does not purport to be a model of a biological brain, we do not need to propose a structure for these perception-units which has any physical components that resemble neurones or synapses. All we require is a functionally equivalent performance. So, to avoid being embroiled in a discussion about neural networks and backward propagation, each perception-unit is envisaged as having the power of a small computer. Each has its own memory store, and its own stored program which attempts to match the pattern of the input signals it receives, with the pattern stored in its own local memory. There is no doubt that that arrangement could be implemented using standard electronic components.

At the very start of this pseudo-evolutionary process we have a very simple mechanism indeed - the equivalent biological organism which is a simple single-celled blob of pond life - one of the flagellates, say. In this organism a light-sensitive dot reacts to sunlight by emitting a transmitter chemical. This diffuses to the flagellum which starts to twirl. If the sunlight is strong the flagellum twirls in one direction. If it is too weak the flagellum goes in the other direction. The organism is, therefore, able to adjust its level in its pond so that it remains in the region where sunlight and food supply are at an optimum. There are no neurones or synapses in an organism like the flagellum although the system as a whole bears some resemblance to a synapse in that it relies on the diffusion of transmitter substances.

In the functionally equivalent simulation, we have a single sensor-unit, two response-units and two perception-units. The sensor-unit sends its signal to both perception-units. Each perception-unit is primed to recognize a signal of a given strength (or above or below that strength). Each, on recognizing the appropriate input signal strength, sends an output signal to the appropriate response unit. If a perception-unit fails to identify the pattern of input which it is programmed to identify, it outputs a negative signal. A negative signal is one which is below some arbitrary threshold level.

The flagellate simulation provides us with our starting point. If we think in schematic terms, we can envisage some simple evolutionary modifications which will multiply the sensor-units, the response-units and the number of perception-units, and diversify their complexity, until we have a very complex system which is capable of dealing with a very wide variety of input signal patterns. We are aiming at a system with many millions or even billions of perception-units and multiple sensor arrays which provide sensory signals from external sources (vision, hearing etc.) and also from internal sources (hunger, thirst, tiredness). It will also have multiple response arrays capable of initiating a wide variety of muscular actions and reorganisations of its own internal structures.

A lot of the work on AI systems has been concerned with devising ways to enable the recognition of in-coming patterns of signals - speech recognition, shape recognition, face recognition, and so on. In the system being described here, some of these problems could be handled by a sufficiently clever hierarchy of complex perception-units (which is not to suggest that the task is an easy one).

In another text (Noble 2005), the author suggested some techniques for shape recognition, which do not involve the use of complex mathematics and are therefore more plausibly the product of gradual evolution. Space does not permit the elaboration of these

methods here. One of the problems we envisage with connection machine recognition systems is that while they are obviously very effective, they appear to require some form of human intervention to teach the system what each pattern means. That is precisely the kind of “solution” we are trying to avoid here (see later, Section 4.2, The Machine Tool Mistake.).

There will be some aspects of these recognition problems which could not be handled at the low level provided by the SRA because they require the use of unseen evidence. A typical example is the recognition of a physical object which is partially obscured. This seems to demand pre-existing knowledge of an object's shape, including parts which may not be visible. A concept (as we shall see at the fourth phase of development) is a compendium of many previous encounters with entities and events of various kinds. Information of that kind cannot be extracted from raw sensory data at a single moment in time, but only from the history of previous experience.

Nevertheless, there is much that the phase-1 SRA could accomplish based on what is immediately accessible in the in-coming stream of sensory perception.

4.2 Issues arising from phase-1

SUBSUMPTION ARCHITECTURE AND THE SRA: The perception-units are semi-independent of each other. In the mechanism proposed here, the only central control linkage between them is a clock mechanism, which outputs a train of pulses. The clock triggers and synchronizes the operation of the perception-units. The output of one perception-unit may trigger a physical response but it may also contribute to the input of another unit. This cannot reasonably be described as a central control, however. It is localized and haphazard. Natural selection will kill off inappropriate connections. In all other respects each perception-unit operates independently of the others and in an opportunistic way. This arrangement is therefore an example of subsumption architecture (Brookes 1990).

SIMULTANEITY: The train of pulses from the clock chops time into intervals. Any two or more signals which arrive at a perception-unit, within such an interval, are considered to be simultaneous.

CLOCKS: A clock mechanism could evolve easily. All that is required is a structure with several perception-units involving delayed negative feedback. Any strongly interconnected system is prone to develop oscillations. Systems engineers often have trouble eliminating unwanted oscillations from control systems of significant complexity.

PRIORITY: Consider what would happen if, in a system like this, two perception-units responded positively and simultaneously to two separate sets of stimuli. The system would have a problem deciding which stimulus, and which response, should take precedence. This would be particularly so if the two responses were mutually contradictory - escape or feed, for example. The decision the system takes would be essentially arbitrary. But given a vast number of similar organisms, which respond differently to a situation like that, natural selection would then favour the assignment of a particular set of priority levels to these perception-units.

VARIABLE PRIORITIES: Priority levels need to be flexible and alterable. Continued stimulation of a particular kind and repetitive responses to that form of stimulation must be able to reduce its priority level, while a denial of service must enhance priority. That means that the system must be able to modify at least some aspects of itself, based on simple observations of its own performance.

RETENTION: Next consider how the system might evolve so that it was able to detect sudden change or movement in the world around it. The ability to do that provides a clear survival benefit. To be able to do that, however, the system must be able to hold on to the condition that pertained during a previous time-interval to enable comparison with the current condition. What is required therefore, is some form of rudimentary memory with a time duration lasting (at least) two clock-ticks. We can call that "retention".

GOAL STATES: Consider also the result of a modification, which triggered a repetition of behaviour in some circumstances. All that is required is the interconnection of two or more perception-units in a loop so that the output of one feeds back to re-stimulate the other (and eventually itself). We would then observe behaviour which we would describe as goal-seeking behaviour. It is more difficult to achieve the converse of that - anti-goal avoidance behaviour but we will consider that issue shortly. The two internal conditions which trigger these forms of behaviour could be called "goal" and "anti-goal" conditions. The mechanism is, of course, entirely automatic and initially arbitrary. No intellectual choice is being made. But again, natural selection will favour the survival, and therefore the continued presence in the gene pool, of goals which confer a survival benefit.

MATURATION: A small adaptation, which would facilitate adaptation to prevailing conditions, could be put into effect if the system was able to count the number of times each perception-unit was able to detect its assigned pattern of signals. After a period of exposure to environmental circumstances, those perception-units which had failed to reach some arbitrary level of use, could be allowed to fade out of existence - thus simplifying the system and freeing resources for other uses. This would constitute a form of learning or maturation which would tune the system to the context in which it found itself.

THE MACHINE TOOL MISTAKE: There is one aspect of this development process, however, which we should avoid introducing. We should never rely on some kind of "background" intelligence or inherent "logic" which would enable the system to "work out" what it needs, when, in fact, such facilities have not yet been developed. That is equivalent to an engineer using machine tools to manufacture the metal components of the very first machine tool. New features of the system must arise naturally, in small steps, from pre-existing features. This point is particularly relevant to the consideration of the evolution of language. We cannot explain the problem of how the mechanism of "deep grammar" could have developed, by relying on some kind of (unexplained) "general intelligence" to bridge an awkward gap in the development path (see later, Section 4.33).

4.3 A summary of the phase-1 characteristics

With an expansion in the range and complexity of the stimuli to which the system can respond, and with the adaptations noted above, the system is ready for the next phase of its development. We should also note that the phase-1 SRA can provide later phases of the system with the following useful facilities and characteristics -

- (1) Facilities for the input of sensory data
- (2) Facilities for the output of control signals to muscles etc.
- (3) An association between input signals and output responses.
- (4) A clock.
- (5) A rudimentary memory mechanism (i.e. retention).
- (6) A set of priorities associated with the perception-units.
- (7) Goals and anti-goal conditions.
- (8) An ability to match patterns of signals.
- (9) Mechanisms for monitoring/modifying its own performance
(i.e. the counting mechanism).

Although the phase-1 SRA may evolve to become a large system with millions if not billions of individual semi-autonomous subsystems, each trying to carry out some task which has been allotted to it by mutation and allowed to prosper by natural selection, it is still (at a detailed level) a very simple mechanism. Even so, as we shall see shortly, it has laid the foundations, which will determine the shape of minds to come.

4.4 A Note on the most Significant Feature of the Model.

Among the mechanisms suggested by this hypothetical model the most significant is the information content which phase-1 (the SRA) makes available to the rest of the system. The behavioural characteristics of the SRA have been programmed by mutation and natural selection, over a very long period of time indeed, to be as they are. The SRA observes the external environment (and its own internal environment) and tries to optimize its internal condition (in terms of goals and anti-goals) in response to those observed changes in external conditions. The other parts of the system, which will be described below, have no additional form of direct access to sensory perceptions, other than what is provided to them by the SRA. The SRA, in effect, becomes a pseudo-environment within which those other, and more "intellectual" parts of the system must operate. Much of the literature on consciousness (from an AI point of view) discounts the possibility of a "Cartesian Theatre" of the mind (Dennett 1991), and that position is not disputed here. But this way in which the SRA provides the rest of the system with an observable environment does constitute a kind of "theatre of the mind". This 'theatre', however, does not consist of images or any other kind of direct analogue representation of external conditions. It is a theatre in which the only players on the stage are the features which have been recognized by the SRA and the actions with which the SRA responds to these recognized conditions. These must be taken together. The SRA is not only serving up the features it has recognized in the external environment but it is tagging each with what amounts to a semantic interpretation of those features. That interpretation is given in terms, which relate to goal and anti-goal conditions, and in terms which can be translated into action only by the SRA.

For example - that large dark shape is not just a large dark shape, it is a shape you should run away from. This is not just a small black speck, of a particular shape, which moves about with a particular pattern of movement, it is something you should try to eat. That thing over there is not just an object, it is fist-sized object and therefore it can be picked up with this kind of muscular action (see Iaconboni 2008). That pattern of sound is not just a sound, it is a cry which triggers your maternal instincts. And so on.

Of course the system, in the early stages of its development has no means to "understand" these associations in conceptual terms. But the raw material for building that understanding is present in the data provided by the SRA.

What that means is that what we may regard as the more "intellectual" aspects of mind, are both constrained, and given a flying start, by perhaps a billion years of evolution. That, however, is not an easy circumstance to emulate if we were to attempt to construct an artificial equivalent system.

For ease of intuitive understanding, we can envisage that the SRA presents all that information (to other parts of the system) in the form of a huge but essentially arbitrary pixellated array, or points of light, each point corresponding to a particular feature or response and each having a certain strength and colour. The metaphor of a pixellated array is being used here only to emphasize the point that there is no special or pre-determined structure to the information provided by the SRA. It evolved without the other parts of the system in mind - since, at that early stage of its evolution, there was no "mind" to have any preconceived ideas.

4.5 PHASE-2: Transient and Persistent Memories

The next stage of development provides the system with a short-term or "transient" memory and a "persistent" or medium-term memory which holds information about selected episodic forms of experience. By storing this information the system provides itself with the ability to augment its genetically programmed responses to various stimuli, by adding learned responses to its repertoire.

In the first instance, to produce the transitory memory, all that is required is an extension of the rudimentary memory store (retention) held by the SRA which enables the SRA to identify movement or change. Progressive increases in the length of time for which a record of a previous condition can be stored brings transitory memory into contention. The survival benefit which this brings, takes the form of an ability to develop longer and more complicated forms of behaviour in response to stimuli which may be very short lived, but which can be preserved in the extended, but still quite short memory trace. For example, an organism with a modest memory could continue to make its escape (or continue to hide) after the stimulus which triggered the escape behaviour had disappeared.

We can envisage this transient memory taking the form of a circular list, which continually overwrites its own tail. We can also envisage a new "state" or condition of the SRA, being added to the list every time the clock ticks (while simultaneously one state disappears from the other end). This transient memory might have a duration of a few seconds.

The advent of memory in that form introduces the problem of memory storage capacity. To extend the size of the memory which can be stored, the system needs to expand the physical brain size. It must also introduce some mechanisms for space economy.

4.6 Issues arising from Phase-2

SELECTIVITY: One method of saving space is to be selective about what is to be preserved. The SRA provides a mechanism for that. If the current state (of the SRA) contains at least one example of a particularly important occurrence (that is, it has a strong reaction from a perception-unit with a particularly high priority level), the relevant information stored in the transient memory could be regarded as being worthy of being stored for a longer period. In that case the entire contents of the transitory memory would be transferred to a longer-term memory store.

EPISODES: If that happens then what would be stored in this more persistent memory would be a set of episodes - each of these would be a chronological sequence of states (conditions of the SRA) with each sequence lasting a few seconds and each leading up to some important circumstance (a goal or anti-goal state). This arrangement could provide an important survival benefit. Since natural selection will have ensured that each goal state is associated with a better chance of survival (and the reverse for an anti-goal state) each episode stored in the persistent memory will provide clues for short-term predictions about what is likely to occur (and a script of how to repeat or avoid the experience). What the system must then do is to try to match the content of the current input stream of sensory experience with some characteristic contents of an episode in the persistent memory. When a match is obtained, the system can then make use of the information in persistent memory, either to repeat the behaviour recorded there (and thus to repeat a goal-state experience) or to avoid repeating the behaviour by overriding the pre-programmed behaviour of the SRA, and (thus avoid re-experiencing an anti-goal state).

Note that since the priority levels associated with perception-units is flexible, the choice of which episodes are selected will depend to some extent upon past experience and the conditions which pertain at the time the selection is made.

4.7 The issue of control.

While the phase-2 part of the composite system is matching current experience with previously stored episodes and choosing a course of action to be followed, the phase-1 part of the system will continue to respond automatically to various stimuli. It must do so because phase-2 can deal only with selected aspects of current experience. But how could the phase-2 system override the automatic and on-going forms of response of the SRA?

Note that the SRA already has a mechanism (in the form of goal and anti-goal states) for triggering repeat behaviour and avoidance behaviour patterns. So when the phase-2 system detects or predicts the advent of a goal or anti-goal condition in the near future, the strategy it could adopt would be to trigger a new goal or anti-goal subsystem within the phase-1 SRA.

Thus the goal and anti-goal states become the means by which the phase-2 system can exercise some measure of control over the SRA. But note that it is not a sure-fire method of control. It can never cover the entire field of sensory experience because the persistent memory, on which it depends to trigger its response to events, is selective. Furthermore, if the SRA is currently experiencing some other high priority condition and behaving accordingly, the pressing of those control buttons might fall on deaf ears. (Forgive the mixed metaphor please.)

4.8 Avoiding Anti-goals

Phase-2 has a particular problem. It can identify a positive script for action likely to lead to a predicted goal-state. But it has difficulty identifying a positive script for action likely to avoid an anti-goal state. In the first instance all it can do (in effect) is to say "whatever you do - don't do this!". That is the equivalent to pressing a "panic button". Until some experience is acquired which demonstrates how an anti-goal state can be avoided, the best phase-2 can achieve is either (i) to induce a fit of hysterics, which may or may not result in a fortuitous avoidance of the anti-goal, or (ii) to recommend a form of exploratory action - small actions which are carefully monitored for the effects they cause.

The SRA, however, does not appear to have any built-in facilities with which it can respond to a panic button, except by making small arbitrary actions. If the system as a whole is to respond in a more sophisticated way to the predicted approach of an anti-goal condition, a more sophisticated addition to the system will be required. And that takes us to phase-3 which will be described in Section 4.10.

4.9 A Summary of the Phase-2 characteristics

- (1) A transient memory (current experience).
- (2) A persistent memory of "important" episodes.
- (3) The ability to match conditions in transient memory with episodes in persistent memory.
- (4) The ability to identify a "script" to be followed in these circumstances.
- (5) Intervention and control of phase-1.
- (6) Directing the focus of attention (for phase-3).

These are the forms of information which phase-2 donates to the system as a whole and to the phase-3 part of the system in particular.

4.10 PHASE-3: Concept Formation.

The next evolutionary phase involves an expansion of the memory, the economy of storage space required for that, and then being able to capitalize on a particular side-effect produced by one particular mechanism of saving space.

4.11 Purging the Persistent Memory

To avoid overburdening itself with the storage of these persistent or episodic memories, the system must purge the memory-store of unused sequences. This requires that the system should have some method of monitoring its own performance. However, we have seen that the SRA is already equipped with a rudimentary version of such a facility by making it possible to count the number of times a perception-unit detects its assigned pattern of signals. In a similar way, the phase-2 persistent memory system needs to count how often each episode within the persistent memory has been utilised. Unused or little used episodes, can then be deleted or just allowed to dwindle and disappear.

4.12 Compression

Another technique which would enable the system to economize in storage space, would be one which is familiar to those working in information communications technology - compression.

To compress the persistent memory, the contents of the memory are scanned, repeating sequences are identified and extracted, the locations from which these repeating chunks of data have been extracted are "book-marked" and the chunks of data extracted are then stored elsewhere (one copy only for each repeating chunk). Let's give the name "compression chunk" to each chunk of extracted material. They have an important role to play in this narrative.

4.13 Concepts

The idea that data compression is a mechanism, which is potentially capable of forming concepts has been around for many years. Various techniques have been tried (see for example Pickett and Oats 2005, Rendall 1985). The appeal of the idea is obvious. Here is a technique which can process a large mass of raw data, extract from it identifiable bundles of data, and do so without human intervention and without relying on "background general knowledge" (the machine tool mistake). The chunks extracted are arbitrary, the only criteria for extraction being that they occur frequently within the raw data. The very fact that they are related to frequently occurring packages of data, however, ensures that they are related to something in the external world which is a significant aspect of human experience.

4.14 The Initial Stage of Concept Formation

The obvious real-life reference targets for these packages or chunks, at the simplest level, are physical objects, or entities. That, indeed, as students of philosophy will know, might be a definition of a physical object - a package of sensory perceptions which occur simultaneously and repeatedly.

It is not that simple of course. A single physical object presents to our senses a package of experience which varies in detail from one occasion to another so it would require a

very clever compression algorithm to recognize the important commonality of those experiences and to discount the trivial differences.

What is being proposed here, however, is that the compression algorithm should be applied, not to raw sense-data, but to data which has already been pre-processed by the phase-1 stage of the system - that is, by the SRA. This was earlier described (metaphorically) as "pixels in a large pixellated array". Within that array each pixel contains information about a pattern of raw sense-data which has been recognized (or not) by a perception-unit. That perception-unit has been programmed to make that recognition. In addition the utility of each perception-unit has already been assured in two ways - by evolution (and hence by natural selection) and by the maturation process which has rejected unused perception-units within the context of the prevailing environment. So the importance and relevance of the features which have been identified, and grouped as a repeating package or chunk by the compression algorithm, have been vouched for. In addition, the inclusion of those response patterns enhances the commonality of some chunks of data which would otherwise be seen as disparate in detail. For example, chunks relating to food items of different sizes and shapes and which could not be easily identified and grouped using physical characteristics like color and shape could still be classified together under the heading "edible". There is experimental evidence which supports that idea (see for example Caramazza et al 1994).

There is another aspect of this idea, which is different from various attempts to use compression as a means to form concepts. In association with the sense-data and the various pre-programmed responses, natural selection has appended to those programmed responses, a measure of their relative importance - those priority levels. Here then is a means by which a heuristic compression algorithm can be directed towards those aspects of the structures which should be matched, and away from others which can be ignored.

4.15 Types of concept

The development of concepts relating to physical objects, is just the first and the easiest type of entity to be identified by this compression procedure. The persistent memory consists of a collection of episodes. Each episode is a chronological sequence of states. Each state is a momentary condition of the SRA (defined in time by the clock-ticks).

There will, therefore, be three broad categories of "chunk" consisting, variously, of one, two or several such "states". These are, (1) physical objects, (2) causal links between states and (3) scenarios, or events.

4.16 Entities

Each one-state chunk will correspond to a physical entity or a condition for which all of the characteristic features occur, or are presented to the senses, within one time-interval.

Some entities are classified, as noted above, not by their physical characteristics - like shape, sub-components, colour etc. - but by the use to which they are put. A representation of use requires the involvement of other physical entities and some desired end-result, or goal-state. That expands the compression chunk beyond the contents of a single state. So while the process of forming a concept associated with a physical entity can start with a

single state, the process does not necessarily stop there. It has to continue and may eventually include a much more complicated representation.

4.17 Scenarios

A chunk consisting of several "states" (more than two) could be described as an event or a scenario (which will normally involve one or more entities). These scenarios are unlikely to be identified by any compression algorithm, no matter how clever it may be, until after the one-state entities within it have been identified and replaced by bookmarks or tags. The compression algorithm could then disregard the differences between what those tags represent, and regard two sequences (or scenarios) as being "the same" if they have a similar skeletal structure (of causal-linkage – see Section 4.18). The tags would then be assigned the role of "actors" within the scenario which need to be instantiated (that is, given their individual identity) when the scenario is used as part of an interpretation (see later).

4.18 Causal-links

These have been placed third in this list because of the peculiarly important role which (it is suggested here) they play in our human psychology and understanding of the world we live in. Chunks, which consist of exactly two chronologically adjacent "states", will be what we might call "a predictive pair". That is, of the two states involved, the first will be a reliable predictor of the second. If the association between them is sufficiently reliable, with the second never occurring without its paired precursor state, such a pairing could be described as a "causal" relationship.

4.19 Causation

The way the proposed system represents causality corresponds to David Hume's analysis of causation. It does not dismiss the possibility that causation is a fundamental component of a mechanism that appears to govern the universe. But it does not endorse that idea either. What it does do, is to represent the formation of a causal-link as something which is perceived and is a fundamental component of our understanding of the universe - that is, it is a fundamental feature of our psychology. With that perspective, and regardless of how the universe behaves, causation can be regarded as an empirical rule which enables us to anticipate likely future experience. At an early and primitive level of the system's evolution this means prediction of what will happen in the next few seconds. But as the system evolves to a more sophisticated level, concepts consisting of causal chains will be formed, stored and then re-used to enable predictions over a much longer time-scale. It enables the planning of future actions.

4.20 Concept development.

Concept formation has to start somewhere and the classification of concepts into entities (one-state), causal-links (two-state) and scenarios (multi-state) refers only to that initial phase of concept formation. Later, the compression algorithm can be re-applied in many different ways, using different "match" criteria.

As simple concepts are formed it becomes easier, on re-application of the compression algorithm, to expand the amount of data included by the repeating detection process. The procedure begins with the foothills and progresses to higher peaks in gradual stages.

An important consequence of re-applying the compression algorithm to concepts which have already been formed, is to produce more and more abstract concepts. This could be achieved by gradually restricting the criteria for a match to be declared. An inheritance hierarchy can be formed in that way, to provide generalized classifications of entities.

Another way in which the nature of entity concepts could be developed would be to widen the search for a match to include contextual information. It would then be possible to recognize a group of entities by virtue of the use to which they are put (tin-openers, weapons), the environmental context of their use (party things) or the psychological effect they have on the people who participate in their use (games) - disciples of Wittgenstein please note.

The same progressive abstraction process can be applied to scenarios. This will be the subject of some discussion later.

4.21 Interpretation.

We now come to a crucial point in the evolution of consciousness. Having established that there is some immediate survival advantage to be gained by the formation of those repeating chunks of data (even one such chunk would confer the slight advantage of a slightly more compact store) we can then anticipate the steady accumulation of more chunks (without having to rely on "miraculous drift"). We also have within the system, procedures that are able to match two chunks of data, or two patterns of signals. Procedures of that kind have been present in the system since the most primitive beginning when the matcher program in each perception-unit was able to match and recognize signal patterns. Also available within the compression procedure is the ability to bookmark.

When there is a need to reconstruct a remembered event or circumstance, the compression chunk, which is stored elsewhere, must be retrieved and restored to its original location. But what is retrieved cannot be the stored chunk itself. That would leave a gap in the store of chunks. So what is retrieved is a copy-version of the stored chunk. Furthermore, as this copy-version is built into its original position, some modifications to it may be required because, while the stored chunk may contain the data which is shared by all occurrences, there will almost certainly be some local variations which have been stored in the tag (or book-mark) ready to be added to the restored chunk.

Here then are the evolutionary precursors of the process, which we can call an "interpretation". What interpretation requires is for those procedural capabilities to be targeted at the in-coming stream of sensory information.

As that data arrives it is stored in the transitory memory. It has been suspected for fifty years or so, that the brain has a working memory space where data can be held and

manipulated in a temporary way (Miller et al 1960) so there is nothing contentious about proposing that the system is able to hold and examine the transitory memory and build alongside it a related structure which uses concepts as construction units.

Particular patterns within that transitory data, which we could call the stream of current experience, can then trigger the retrieval of the copy-versions of concepts, which have the same characteristic patterns. Those copy-version concepts can then be placed in working memory and linked together using the memory reconstruction procedure. The same procedure will be used when the system deals with spoken language, so a discussion of the interpretation mechanism will be deferred until that point.

4.22 Indexing the concepts

Fast access and retrieval of the stored concepts is required if the system is to be able to interpret events in real-time. The way an IT system usually achieves fast retrieval from an online archive is by indexing the stored items. There seems no reason why the brain-model could not use the same technique. This would require only a small modification of the existing matching routines (used by the perception-units) and a small adaptation of the concept creation mechanism, to identify features of each concept which are not widely shared by other concepts, and the storage of these, together with a pointer facility already developed for the book-marking of concept-compression process. Small steps. Each a modification of some pre-existing feature of the system.

4.23 Dream-sleep

If the usage of the concepts is, of necessity, a very fast real-time procedure, the indexing of the concepts is, in contrast, a very slow procedure. It could not easily be carried out while fresh input is being processed. The compression mechanism, together with the other forms of memory processing, make it necessary, therefore, for the system to have regular quiescent periods during which the compression processing, concept formation and purging of unused memory entries, can take place without the interference created by new sensory experience. This, it is suggested, is the basis for dream-sleep.

During this indexing procedure the system will find itself dealing with a mixed collection of index-tags drawn from several different concepts. If the interpretation process was then to be invoked (that is, if, with a view to discovering an appropriate location for the storage of a new item of information, there is an attempt to fit a "picture" together using the pieces available) a temporary, ad hoc "puzzle" picture could then be constructed from this mixed bag of index-pieces containing elements of all those different original concepts - some new, some very old. The result produced by that would be likely to be very odd indeed and would often reflect the presence of particularly traumatic past experiences (or events) mixed in with new material.

4.24 A Summary of the Phase-3 characteristics

- (1) Compression of memory storage
- (2) A collection of compression chunks which are the basis for concept formation.
- (3) Initially those concepts are of three types - entity, causal-link and scenario.
- (4) The ability to construct an interpretation of a subset of events as they are experienced.
- (5) Causal-links which have particular significance for the prediction of future events.
- (6) Re-application of the compression algorithm to produce more generalized forms of each type of concept (entity hierarchies and abstract scenarios).

4.25 PHASE-4: MIND and SELFMIND

At this stage, the brain-model has developed the ability to interpret on-going events and, from that interpretation, to predict future events based on past experience. In that way it can enhance its chances of achieving goal-states and avoiding anti-goal states. It is a sad fact, however, that of all the things around us, the most dangerous, as well as being the greatest source of succour, is the behaviour of other human beings. It is, therefore, important that the brain-model should develop an ability to second-guess the behaviour of other people.

4.26 Predicting behaviour

Initially, this ability will depend upon the reading of overt clues - posture, hand gestures, tone of voice, facial expressions and eye-pointing, are the obvious examples. If observations of these clues occur in the persistent memory, the compression algorithm would be able to detect a repeating pattern of behaviour for which groups of these overt clues are predictors. Thus one group of clues will indicate aggressive behaviour and another will predict gentleness. The result of compression applied to these observations will be abstract concepts which could be labelled "anger", "kindliness" and so on. At this stage of course, the system has no words with which to refer to these abstract concepts and there will be no need for the system to be conscious of having developed the concepts. But that does not prevent or obviate the need for their formation. The next stage is the identification of a two-state concept (or causal-link) which links those abstract concepts with their predictable consequences -

Anger => aggressive behaviour
Kindliness => gentle behaviour
Joy => joyful behaviour
Sadness => tearful behaviour
etc.

4.27 Theory of MIND

The system is now in a position to predict future behaviour by observing those overt clues, classifying them as what may be called "a state of mind" and from that, predicting likely behaviour. The concept of a state of mind is therefore nothing more than an abstract classification of various overt clues.

A further abstraction, which could be developed from all those different types of states of mind, provides the more general abstraction MIND. The concept of MIND is therefore the causal precursor of deliberate behaviour. And that is the explanation of the difference between deliberate and accidental behaviour and a clue as to how we can represent that difference in the hypothetical brain-model. The representation of deliberate behaviour will contain a representation of the action, a second representation of the action which is stored inside the MIND structure of the person concerned, and that representation will show that the action leads (in the person's mind) to a goal-state (of that person). That point deserves some emphasis.

According to this model, concepts are formed from the various conditions present in the phase-1 mechanism of the SRA. That means that the conditions we have called "goal" and "anti-goal" states will become established within that set of concepts. And that enables the explicit representation of elusive concepts such as "desire". To achieve that, the mechanism constructs the representation of a scenario (or a causal chain linking several other concepts and conditions), which leads up to some goal-state. That scenario is then placed inside, or attached to, the MIND of the person who has that desire. The assumption is that a person (or an animal) can always be expected to try to achieve a goal-state. However, that technique also permits the representation of several conflicting goal-states, some short term (and perhaps "selfish"), while others may be long term (and perhaps "altruistic"). The representation of "morality" then becomes the representation of behaviour which strives for the long-term, and unselfish, and avoids short-term gratification.

Note that because the compression algorithm can be applied several times to the basic observable clues to behaviour (and to the results produced by that compression) there will also be several levels of sophistication and abstraction which could be described as a "Theory of Mind". We do not therefore envisage some sharp division in this respect between humans and other species. The possession of a Theory of Mind (or ToM) is a mental characteristic which can be acquired gradually in small evolutionary steps.

4.28 Theory of SELFMIND.

There is not much survival benefit to be gained from an ability to predict future events and the likely behaviour of other people, if the system is not able to predict its own likely response to those future events. There is a tendency to assume that any intelligent brain mechanism will automatically know what it itself is doing and how it will behave in the future. But to make that assumption is to fall into the machine tool mistake. We must ask how it can know these things. What is the mechanism?

The primitive SRA is able to receive signals from internal sensory arrays which will detect internal conditions like hunger, thirst, muscle fatigue, alarm, disposition of limbs, interrupt signals (pain), and similar. Compression and abstraction of these will result in a

generalized concept SELF which is a representation of the system's own physical body. Add to that the ability to detect the system's own condition with respect to goal and anti-goal states and the external observation of how the body behaves in response to those internal conditions. All of these sensory signals are placed in the transitory memory, and if they are selected because they include some important condition, they will be stored in the persistent memory. And that means that these sensory signals and their associated responses are available for analysis by the concept forming process of compression.

We then have the data necessary for the formation of a Theory of SELFMIND concept. This is the abstract form which groups all of these overt clues (including those overt internal clues) and is the notional causal precursor of the system's own behaviour. In this model, much of the mechanism is automatic and beyond the control of the upper stages of its mental development.

4.29 Consciousness and Conceptual Memory

That SELFMIND structure becomes the repository for the interpretation structure. This creates yet another memory store. The interpretation structure is constructed from concepts and this new kind of memory is a chronological sequence of interpretation structures, segmented by switches of attention. Every object within the interpretation structure is not simply an object as it is observed at a particular moment but is a compendium of that and all previous encounters with the object. The system explicitly represents itself as interpreting its own experiences and it remembers those experiences. It represents itself as predicting its own future behaviour. It represents itself as representing itself.

It is with the advent of this complex recursive self-representation and self-prediction mechanism that the brain-model becomes fully conscious - because that is what being conscious means - doing the experience of representing an experience and "knowing" - that is, having a conceptual analysis of events - what those experiences imply by way of consequent future experiences. Being conscious means being the procedure which does that.

Note, however, that there is no single threshold over which the system must step in order to convert itself from a completely unconscious mechanism into a completely conscious one. Consciousness, yet again, is a procedural condition which is acquired in gradual stages. A significant part of the early stages of that gradual approach is that part of the system which could be regarded as being "unconscious". We may conclude, somewhat paradoxically, that having an "unconscious" part of one's brain is a pre-requisite for being "conscious". If that is the case, however, the prospect of finding the "neural correlates of consciousness" must be somewhat reduced.

4.30 A summary of the Phase-4 characteristics

(1) The re-application of compression procedures to produce abstract concepts of various kinds -

- (a) The classification of entities in a structured hierarchy.
- (b) The identification of the generalized concept - "cause".

(c) Abstract scenarios like "justice" and "duty" within which the individual details of the various contributing events are ignored and the only thing which remains are the skeletal causal structures and the mental attitudes and intentions of the actors within the scenario.

(2) A Theory of MIND

(3) A Theory of SELFMIND

(4) A memory store of the on-going interpretation of experience within the SELFMIND.

4.31 PHASE-5: Language

The fifth and last stage of development of this model is the acquisition of a language facility. According to the thesis offered here, a language faculty is not an essential component of conscious, but it does help to raise consciousness to higher levels, particularly with respect to the development and use of abstract concepts. As a fortunate side-effect it also provides us with a logical proof of robotic consciousness (Section 7).

4.32 The Implausibility of Deep Grammar as an evolutionary precursor

Philosophers and linguists have long debated the relationship between consciousness and the ability to use spoken language. Some have suggested that language is an essential precursor to being conscious and/or that it is impossible to form concepts without having a facility for language (Jaynes 1976). In recent years, however, as animal psychologists and behaviourists have studied chimps, dolphins and the like, it has become clear that these animals are much more able to utilise various forms of linguistic communication than we had previously supposed (Fasold 2006). These observations suggest that the Chomskian idea that grammar (or a deep universal grammar) is an essential requisite for the use of a language (Chomsky 1971), is implausible (Jackendoff 2009).

4.33 Pinker's Defence

Stephen Pinker has tried to defend the idea that deep grammar is an evolutionary precursor to language. He is not alone. But his writings on this topic has been influential so we shall take them as typical of the genre.

Pinker deploys a number of arguments. Some of these are specific and detailed and others are more general. Consider first, his claim that in a statement such as "*it is raining*" the word "*it*" has no meaning, no semantic interpretation, but is instead merely a syntactical device or "place holder". This dependence upon syntax for a classification, he claims, proves that syntax has a basic role in the language faculty (Pinker 1994, p42).

However, the classification of "*it*" as a place-holder, depends upon the presupposition that syntax has the basic role which Pinker claims it has. The argument, therefore, is an exercise in circular logic.

In the model of language being offered in this text, "*it*" does have a semantic interpretation. That pronoun, in that context, refers to a general contextual environment or ambience and switches attention to that context. It is analogous to a reference to an

environment variable in a computer program. In terms of the kind of interpretation procedure described earlier, "it" tells the listener to attach the representation of "raining" to that general ambiance and not to any particular constituent part of the environment.

Next, consider Pinker's argument about evolution.

"To be fair," Pinker writes, "*there are genuine problems in reconstructing how the language faculty might have evolved by natural selection ... If language evolved gradually, there must have been a sequence of intermediate forms, each useful to its possessor ...*" (Pinker 199, p365).

Indeed.

"Who," he asks, "*did the first mutant talk to?*" (ibid). He answers that pertinent question in two ways. He supposes that the first linguistically gifted individual might have talked to "*the fifty percent of brothers and sisters, sons and daughters who shared the new gene by common inheritance*". This appears to presuppose that there is just a single new gene involved (in contradiction to the first of the two quotations). It also implies that one of the parents of that first gifted individual was also in possession of that special gene. So that first gifted individual was not really the first after all. The parent, who was the first, had to keep this amazing new faculty in abeyance for many years until those sons and daughters came along. Not much survival advantage to be gained then in those early years.

But Pinker goes further. In a second attempt, to explain that anomalous evolutionary development, he suggests that the first linguistically gifted individual could have conversed with others who did not have the "deep grammar" gene, but who would have been able to understand this special individual "*just using overall intelligence*". Ah Ha! The machine tool mistake! There must after all be a way of understanding language which does not depend upon a genetically endowed knowledge of deep grammar.

Two obvious questions - (1) how does this "*overall intelligence*" language ability evolve and (2) why cannot it evolve further to become a more sophisticated language facility?

We need an alternative way in which language could develop and one which is much more plausible than the "deep grammar" idea from an evolutionary point of view. This paper does that, and offers answers to both these questions.

4.34 The Construction-kit Theory of Language

The suggestion is that all the higher mammals, which can reasonably be supposed to be conscious of their own behaviour, will also have the mental mechanisms needed for a rudimentary form of linguistic communication. It does not need to be a spoken language. The essentials of that mechanism are -

(1) the ability to analyse the transitory memory for the presence of signal patterns which are characteristic of certain concepts. These signal patterns can come in various forms - gestures, facial expressions, etc. and also in the form of sounds which might be spoken words,

(2) the ability to form, store and retrieve these concepts,

(3) the ability to put them together to form an interpretation structure, and

(4) the storage of that interpretation in a SELFMIND.

Most animals will not be able to do all of that to any great extent. Most will deal only with one concept at a time. But some may be able to put two or even more concepts together to form a new composite interpretation of events. It has been reported for example that Washoe, the female chimp raised by the animal psychologists as though she was a human child, and who learned American sign language, was able to construct the double-sign for "water bird" when she first encountered a swan [New York Times 11/1/7].

To provide the full mechanism of communication, a creature also requires the ability to put that mechanism into reverse - to analyse an interpretation structure into its component concepts, to identify the symbols, tokens words or whatever, which serve as labels for these concepts, and then to utter those symbols in whichever way the creature's anatomical structure permits.

This mechanism (in different forms) has been called variously "The Construction Theory of Language" (Noble 1988. Noble 2005) and "Conceptual Semantics" (Jackendoff 2009).

At a primitive level, communication is restricted to single word utterances or to simple two and three word groups. But humans are able to form more complicated concepts, including meta-concepts (see later). With these meta-concepts humans can overcome the problems which arise when an essentially linear form of communication is used to describe circumstances which are four dimensional (the three spatial dimensions and the time dimension).

4.35 Objections

This notion, the "constructive" or "combinatorial" properties of language and how it operates has often been suggested in the past and just as often been discounted. The main arguments against it raise two issues -

- (a) How do questions and commands fit into this picture?
and
- (b) What kind of concept structure could be associated with a word like "how"?

Both objections can be answered easily once we have accepted the idea that the system can construct a representation of MIND and its contents.

4.36 A Response to Objection (a)

A declarative sentence describes a condition of the world ("The door is open").

A query ("Is the door shut?") also describes a condition of the world and, in particular, the internal mental world of the speaker. Within that mental world there are two hypothetical conditions (The door is shut and the door is open). The query also describes an additional mental condition of the speaker such that a request is being made for help in resolving the ambiguity.

A command ("Shut the door!") also refers to an internal world model. It tells us, in this example, that the door is open. It also tells us of the speaker's mental world in which the door being shut is the desired condition. We could also read into the statement the implication that the speaker will be annoyed if the listener does not comply.

If we extend the notion of a world description to include mental worlds in that way we can see that all sentence forms can be interpreted as world descriptions.

4.37 A Response to Objection (b)

The construction kit theory of language resolves the problem of what concepts we should associate with words like "how" or "because" or any of a host of similar terms. An analogy might help. Consider one of those modelling kits which enable enthusiasts to assemble plastic models of aeroplanes and battleships. Most of the kit will consist of plastic components which represent parts of the finished model - wings, tail-fin, engine, and so on. But the kit will also contain two additional components - a set of instructions and a tube of glue. If the construction theory of language is correct, then each uttered sentence should contain analogous components. The explanation for the role played by those additional words like "how" and "because" is that they are equivalent to the instructions and the glue. They indicate the point of attachment of other words in the sentence and they provide the glue which holds them to their appropriate attachment points. The word "because", for example, indicates that the most appropriate point of attachment will be concerned with causation. It introduces an extra causal link structure and in effect tells us that *this part* of the interpretation structure is the *cause of that other part*. The default interpretation of a word like "on" will indicate a point of attachment concerned with spatial relationships, but may have a more general interpretation which requires the properties of the two structures being linked, to be examined and matched as far as possible.

Example – "He did it on a cold winter's day". One part of this utterance refers to *an action*. The other part identifies *a time*, so the complete interpretation identifies the time of the action.

We can classify the meaning of words like "how", "on" and "because" as "meta-concepts" (concepts about concepts). The ability to form meta-concepts is almost certainly confined to the human species.

The order of the words in a sentence provides some information about the chronology of the action described and the proper points of connection. We can see therefore, that grammatical structure has not been rejected totally. It creeps back into this narrative, but as a secondary refinement to the language facility, not as its foundation or as a requisite precursor.

Perhaps the most radical feature of the construction theory, is the role played by causal links. These are represented as structural elements explicitly, in their own right, and not merely as pointers inserted into other structures. Causal links are, therefore, able to enter into structural relationships between other components of a representation, including other causal links. This enables structures of baroque complexity to be built. It is possible, for example, to represent the concept of "permission" as "a cause which causes a cause", and "prevention" as "a cause which causes the negation of a cause".

4.38 Intuition (and the role of recursion)

An objection which is raised frequently to any suggestion that consciousness has a simple physical explanation, is that that idea contradicts our intuitive subjective feelings about consciousness (Searle 1993). For some, their own intuition provides them with irrefutable evidence that something very strange and inexplicable is going on in our brains. To be intuitively satisfactory, therefore, any physical explanation of consciousness must also explain the mechanism of intuition and explain why it is that that feeling of mind-body separation is so common.

Consider again the recursive structure of the mental representation which the system must form. Self-representation always threatens infinite recursion. Some levels of recursion can be accommodated, but at some point, to avoid a computational impasse, recursion must be brought to a halt. The obvious solution to this problem is the progressive degradation of the representation. At some point the self-representation will then become a representation of a simple physical object with no inner representation of itself. At that point recursion will stop.

Some aspects of the brain mechanism must, therefore, be omitted from the self-representation. Those omitted components, however, will still exist in reality, will still continue to operate and must therefore continue, in reality, to make a contribution to the full procedure. But so far as the self-representation is concerned, the information provided by those unrepresented components must appear to come from nowhere - or from some aspect of mind which hovers behind or above the representation of SELF. Here then is an explanation of the what Gilbert Ryle called "the ghost in the machine" - that mysterious SELF which seems to hover about in our brains and gives rise to our intuitive feeling of SELF.

4.39 Qualia

The brain model offered here provides a simple physical explanation for "qualia" - that is, for the subjective experience of pain, and the seeing of red, and similar. It suggests that to "have a quale" is to perform a procedure which processes and records the receipt of sensory information of some kind. But the issue of qualia is so central to the case against a materialist explanation of consciousness, and so carefully enveloped in mystery, that no account of consciousness can be complete without dealing with it explicitly, using the terms favoured by those anti-materialists.

And therein lies a difficulty. If we adopt the terminology of the anti-materialist we are in danger of being trapped in an inappropriate bystander viewpoint, which is their favoured position. It is then tempting to make a comment such as "*The experience of colour or the feeling of pain is just what it feels like to be performing the procedure described.*"

That would be a mistake because the words underlined "what it feels like" reintroduce the idea of a "feeling" which can be "like" something else. Like a stage conjuror inducing us to select a the card he is pressing on to us, that choice of words presses on to us the notion that a "feeling" is some kind of epi-phenomenon which can be taken for granted, and as a whole, without the need to explain the mechanism of this feeling. If performing the procedure is described as being "like" something, then we have to ask who or what it is

that is having that experience and what is happening inside *that other* mind – not the one we thought we were talking about.

We must choose our words carefully. Pain is the performance of a particular procedure. We do not *have* an experience *while* the procedure is being performed. That choice of words once again lifts the experience to one side, and presents it as an inexplicable whole which we then *have*. That is not the way it is. We *do* an experience. *Having* an experience describes the situation when it is viewed from the outside. *Doing* a procedure describes the situation when viewed from the inside. It is the same procedure in both cases. If we mix the descriptions so that we are both *doing* the procedure and *having* the experience at the same time then *EITHER* there must be two procedures, one described and the other taken for granted, *OR* we appear to be both inside and outside a single mechanism at the same time.

Is that possible?

Well ... actually ... it can appear to be possible. That is exactly the situation which the self-representing mind creates. The mechanism executes the procedure and as it does so it is also constructing a self-representation of itself doing the procedure. So we are inside one procedure and observing a representation of that same procedure from the outside. That, perhaps, is the source of the difficulty which many people experience when they try to understand the physical explanation.

5. Implementation

The description of the brain-model made use only of computational ideas, which could be realized using conventional electronic components. Implementation of the whole system is, therefore, theoretically possible. However, the practical difficulties of doing so are very great indeed.

5.1 The Problem of Scale

The first of these difficulties is the scale required. The manufacture of one perception-unit would be easy. The manufacture of several billion is a different matter, especially since a large proportion of them would not be identical.

Two methods might help. The first is to rely on evolutionary programming. One particular perception-unit could be constructed, provided with a hand-crafted matcher program and with a carefully specified pattern of signals which it is required to recognize. That design could then be replicated a large number of times after which the population of identical perception-units would be subjected to small random variations. The mixed population, which that produced, would then be confronted by a realistic natural environment and the next generation of units would be "bred" from the subset which were able to recognize their designated signal pattern most readily.

An alternative approach would be to start with the creation of a population of perception-units which have a random assignment of programmes and specified patterns. That approach is likely to have a very large proportion of rejects, but it is also likely to throw up a few specifications which would have been missed by the first method.

Either way, however, the development of the SRA is likely to be a very slow process indeed and one which requires a huge investment in resources.

5.2 The Problem of Raw Materials

And that is only the first phase of the development process. The later stages cannot even begin until that first phase has been to a large extent completed, for it is the first phase which produces the collection of data which becomes the raw material on which the later phases depend. It is rather as though we were trying to build a house but found that there were no existing clay mines or factories which would produce the bricks from which the house would be constructed. Various infrastructure projects – such as clay-mines, factories, lorries, roads to bring the materials to the site and so on - would need to be put in hand even before the foundations were poured or the first brick was laid.

5.3 The Problem of Maturation.

It is also difficult to see how that initial phase could be speeded up since progress is dependent upon the chance encounters of the developing system with various features of that naturalistic environment. The only guide we have to the time required by this process comes in two parts - (i) the several billion years or so that it took for real evolution to produce the first organisms with the capability of an insect or a mollusk, and (ii) the ten to twenty years it takes for one human child to acquire mental maturity.

Various subsystems within the specification of the hypothetical brain mechanism could be examined for feasibility in the laboratory. The construction theory of language could be tested using simplified manually constructed concepts in association with a small vocabulary. In an earlier book, this point was illustrated using some tentative concept structures (Noble 1988).

5.4 Embedded Functions

One idea contained in that book, was that the software designed to carry out the interpretation procedure should not be contained in a single monolithic program. It should instead be split up into small routines, each designed to do a very particular task - like finding within an utterance the most appropriate causative agent for an action specified by a given verb. A special subroutine of that kind would then be embedded in the concept structure associated with that verb. The routine would only become active if that verb was actually used in an utterance. Its function would be to scan the specification of the likely agent and then search for some referent within the sentence, which matched that description. Once again we see that the process involves a matching algorithm. The resulting process is analogous to the way “methods” are encapsulated within “objects”. There is very little in the way of a central control procedure. Instead the concepts themselves reach out to find the most appropriate point of attachment for themselves, and they can do that in parallel. It means that the construction of the interpretation structure

would be a process that is almost organic. The bits come together seemingly of their own accord.

But the details of the language system being proposed here, are relatively unimportant. They acquire importance only in so far as they demonstrate that the general principle is feasible. It is that general principle that is important – that a complex and sophisticated facility for the use of language can evolve from a much simpler and pre-linguistic condition concerned with the formation of concepts. These concepts would be primarily used for the interpretation of sensory – that is, non-linguistic information. An innate knowledge of some special facility like a “deep grammar” is not a requirement.

6. So what must a Brain DO to be Conscious?

Note the emphasis on the word “**DO**”. That is the first and most important feature of this explanation of consciousness. Consciousness is not some additional magical epiphenomenal property which a brain procedure can possess or create as a side-effect. It is not even an un-magical property. It is something a brain procedure can do. According to the hypothetical model described here, that activity has the following components –

- (1) A mechanism of sensory perception targeted at internal as well as external conditions.
- (2) A means to respond to those sensory perceptions in terms of muscular actions and also in re-organization of its own mental structures.
- (3) A stimulus-response automaton (SRA) which links those in-coming signals with those responses, identifies goal-states, and appends priority levels.
- (4) The ability to record the present and past performance of the SRA in a way that permits those stored records to be analysed.
- (5) The ability to form concepts based on its analysis of those records.
- (6) The ability to re-process concepts to form more generalised and more abstract concepts.
- (7) The ability to recognise the presence of those concepts within the on-going stream of sensory input and the construction of an interpretation of that experience in terms of concepts (and its own goal-states).
- (8) The ability to make predictions about future events (and its own likely responses to those predicted events) based on its interpretation of current events.
- (9) The application of those concept-forming procedures, to the overt clues of animate behaviour, to form the abstract concept we call “a Theory of Mind”.
- (10) The application of those same procedures to overt clues concerning its own behaviour, augmented by those arising from its own internal conditions, to form a “Theory of SELFMIND”.
- (11) The ability to store its interpretation of events within the SELFMIND structure
- (12) The ability to avoid infinite recursion by NOT recording in that SELFMIND some selected aspects of its own brain procedures.

These are the requirements which emerge from the hypothetical brain-model described here. It is quite a long list. Even so, we repeated the caveat given in the introduction – this model does not purport to capture the full complexity of a real biological brain. It is intended only to set out the absolute minimum needed for consciousness.

It is possible that some other brain-model may be able to achieve the same overall result. It is suggested, however, that though there may be differences in detail, the same general pattern will be present. That is, there will be a number of features, all of which will be useful in themselves, and none of which will represent some sharp dividing line between unconscious and consciousness. Some of these component features while being essential for consciousness will also be present in mechanisms which would not be described as being conscious. Awareness, in the form we have called here "instantaneous awareness", will be present from the start. That is nothing more than an ability to sense the external world (and the internal world) and to react to that in a way that increases the organism's chances of experiencing a goal-state (and therefore its own survival chances).

That primitive ability will be augmented gradually, first by preserving sensory information for a short time in memory, then by extending the duration of that memory, then by remembering selectively only those aspects which are association with high priority events, and then by using causal-linkage to project that understanding of the past and the present, into the future. All this will be augmented further by interpreting events (in conceptual terms) and by being aware of that interpretation.

Every concept, because it is a compendium of information gained from several past encounters with some aspect of life, contains an element of information we might describe as a representation of "what comes next". This action, which will include the storage of that interpretation in SELFMIND (and an instantaneous awareness of that), will represent the state of affairs we call "consciousness". Conscious, then, is an awareness of remembered awareness, the remembering of that, and the knowing of what will come next. The simple unconscious organism has instantaneous awareness but can live only in the immediate present. A conscious organism (at the other end of a long development process), lives in the past, in the present and in the future. It is also aware of its own presence within that context, and can control its own ability to modify it. To be conscious is to be able to predict one's own likely response to predicted events. To be conscious is to construct an explicit representation of one's own awareness.

We repeat a comment made earlier. Being in part unconscious is an essential requirement for being conscious.

7. A Final Word - A proof of robotic consciousness.

Suppose that we were able to overcome all the practical problems of implementation and were able to produce a robotic system, which could perform all the procedures specified in the hypothetical brain model offered here.

What then?

This robot would possess a set of concepts which would include the concept of pain, of redness, and of all the other things to which the enthusiasts for supernatural or epiphenomenal "explanations" of consciousness draw our attention and classify as qualia. The robot must have these concepts because it will have encountered these aspects of life and stored the sensory experience in memory and then processed those memories to create the relevant concepts. As we have frequently remarked, a concept (as envisaged here) is a compendium of past experience. If this robot was then able to discuss these concepts "he" (allow that pronoun please) would need to have a structure which would act as a reference for each word, including elusive ones like "pain" and "redness". The system proposed here

for handling language would not work if the robot did not have these references. Note that it is not necessary that the robot can articulate the meanings for these various words. But it is necessary that he has them and that his interpretation mechanism can manipulate them.

So what would those concept-structures contain in the way of information? For "redness" the robot would have various records of patterns of signals and patterns of responses generated when it received electromagnetic radiation in the appropriate range of wavelengths. Particular examples would be related to particular entities and events - sunsets, post-boxes, bricks, red hair, and so on. Stored in those concepts would be references to the emotions associated with those experiences (goal and anti-goal state conditions). "Redness" would be the general classification for all of these formed by compressing and abstracting them. It would therefore contain all of those potential references too. The concept "pain" would contain references to particular experiences when an alarm-signal interrupted all other mental procedures, when attention was focused on immediate action required to avoid continuation or repetition of the experience and when these experiences triggered the appropriate responses.

Even if we do not accept that the interpretations which the robot would be able to form by using those concepts, were equivalent to our own, we have to ask how adequate or inadequate they would seem to the robot himself.

He has no other version of these things with which to compare them. So they would seem to him to be completely adequate. The trouble which many people have in accepting this account of things is due to the way they apply their own judgements (which they assume to be perfectly informed) to the robot's view of his own understanding - *He thinks he is telling the truth but we know that he is not*. That is not a valid stance. If a person (or a robot) thinks he is telling the truth then he not guilty of deception even if we disagree with his account.

The problem of assessing the apparent "reality" of his understanding - applies particularly to his understanding of the word "*consciousness*". How would he interpret it? And what would be the content he would associate with consciousness? Clearly he must include within that concept, references to all those experiences relating to redness and to pain - and to a host of other so-called "qualia". So if we asked him to tell us if he feels that he himself, is genuinely conscious, what would be his answer? In so far as he understands the term, would he honestly regard his answer as the truth?

Before answering that question, please reflect upon the criteria which we use to decide (and to discuss) our own conscious experience. Is it not the case that consciousness is a self-validating phenomenon? If you *think* you are conscious, is it not the case that you *are* conscious? How could we ever have a circumstance where a person thought he or she was conscious but was not actually conscious? Even if you are dreaming (and within that dream you believe yourself to be conscious), is it not the case that you are conscious - within the context of that dream?

If self-validation is valid for the judgements we make about ourselves, what justification would we have for rejecting the robot's own judgement that he is indeed conscious? Do we have some other criterion which we could use to judge the merits of his claim, that applies only to robots and not to humans? Do we really think it is possible to have an unconscious robot who honestly believes that he is conscious?

References

- Brooks, Rodney (1990) Elephants don't play chess. *Robotics and autonomous systems* 6, (1990) 3-15
- Caramazza, Alfonso, Hillis Argye, Leek Elwyn and Miozzo Michele. The organization of lexical knowledge in the brain: evidence from category- and modality-specific deficits. In *Mapping the Mind*, (eds: Hirschfeld L and Gelman S) Cambridge University press 1994.
- Chomsky, Noam (1971) *Syntactic Structures*, Mouton.
- Dennett, Daniel (1991) *Consciousness Explained*, Little Brown & Co 1991
- Fasold, Ralf Connor-Linton, Jeffrey (2006) *An Introduction to Language and Linguistics*. Cambridge University Press.
- Hare, Brian Call, Joseph Agnetta, Bryan and Tomasello, Michael (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour* Vol 59, issue 4 pp771-785
- Humphrey, Nicholas (2008) *Getting the Measure of Consciousness*. *Progress of Theoretical Physics Supplement* No 173.
- Iaconboni, Marco (2008) *Mirroring People*, Douglas and Macintyre Ltd.
- Jackendoff, Ray (2007) *Language, Consciousness, Culture*. MIT Press.
- Jaynes, Julian (1976) *Origin of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin Co.
- Koza, John (1990) Evolution of subsumption using genetic programming. in F. Varela P. Bourgine (eds) *Proc. First Eur. Conf on Artificial Life*. Camb Mass. pp110-119 MIT Press.
- McCarthy, John (1996) *Making Robots Conscious of their Mental States*, *Machine Intelligence* 15, OUP
- McDermott, Drew (2001) *Mind and Machine*, MIT Press
- Miller G.A. Galanter E and Pribram K.H (1960) *Plans and the Structure of Behaviour*, Holt Rinehart and Winston, New York.
- Minsky, Marvin (1985) *The Society of Mind*, Simon and Schuster Inc.
- Noble, Hugh (1988) *Natural Language processing*, Blackwell Scientific Publications. (available online as a e-book from www.tartanhen.co.uk > books)
- Noble, Hugh (2005) *Operational Consciousness*, Tartan Hen Publications (available online as an e-book from www.tartanhen.co.uk > books)

NOBLE

Penn, Derek and Povinelli, Daniel (2007) On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind' *Philos Trans R Soc Lond B Biol* 2007 v.362 731-744

Penn, Derek, Holyoak Keith, and Povinelli, Daniel Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioural and Brain Sciences* (2008) 31, 109-178

Pickett, Marc and Oats, Tim (2005) The Cruncher Concept formation using minimum description length. *Lecture notes in computer science* 2005, 282-289,

Pinker, Steven (1994) *The Language Instinct*. William Morrow and Co.

Rendall, Larry (1985) Substantial Constructive Induction using Layered Information Compression Tractable Feature Formation in Search. *IJCAI* 1985, 650-658.

Searle, John (1993) The Problem of Consciousness. *Social research* 60 (1) 3-16.
also <http://www.ecs.soton.ac.uk/~harnad/Papers/Py104/searle.prob.html>

Shapiro, Paul (2007) Moral Agency in Other Animals, *Theoretical Medicine and Bioethics*. Vol 27, Number 4 357-373

Sloman Aaron (1994) Supervenience and Causation in Virtual Machinery. Work in progress <http://www.cs.bham.ac.uk/research/projects/cogaff/talks> (updated Jan 31st 2011)

Togelius, Julian (2004) Evolution of subsumption architecture neurocontroller. *A.I.* vol 15 No1/2004