Architecture for a Cognitive Conscious Machine

Proposal White Paper

Rodney M. Goodman rodgoodman@gaeacorporation.com
Gaea Corporation (626) 419 4920
http://www.rodgoodman.ws/

In Collaboration with:
Igor Aleksander, Imperial College, U.K.
Owen Holland, University of Essex, U.K.
Christof Koch, California Institute of Technology

Objectives:

This proposal describes a research program to develop the technology for constructing, programming, and training a cognitive conscious machine – a computational device which is conscious "in a similar way to ourselves" and which therefore possesses the characteristics of conscious human cognitive intelligence unattainable using current approaches. The key enabling idea in our approach is an architecture that features understandable internal models [Holland & Goodman 2004]. This work falls in the area of "Cognitive Computing"; it is revolutionary in that we propose a paradigm shift in the design and architecture of intelligent machines, which will, if successful, underpin a revolutionary approach to many aspects of computation itself. It also has the potential to revolutionize the relationship between humans and computers; in particular we expect this research to result in significant advances in the development of human machine interfaces, making intelligent interaction with everyday machines such as cell phones, cars, etc possible.

Motivation:

The original and ultimate aim of artificial intelligence was, and is, to develop systems that exhibit human-like intelligence. Such systems, if they existed, would be so different from, and superior to, current AI systems that they would effectively constitute a revolution in information technology. We believe that consciousness is an essential component of any truly human-like intelligence, and that true artificial intelligence will not be achieved unless and until it incorporates some adequate form of artificial consciousness. We therefore propose to directly address the problem of developing artificial consciousness, by constructing a series of biologically inspired physical systems based on current neuroscientific knowledge of the structures supporting consciousness, and studying and demonstrating the emergence of key features of consciousness under the appropriate environmental circumstances. We believe that the structures revealed and developed by this approach will enable the incorporation of some form of artificial consciousness into purely computational artificial intelligence systems.

Previous work by Aleksander (1999) using computer simulation has resulted in the development of neural network architectures which, through a technique of iconic learning, can display forms of artificial awareness, and the capability of visualization of internal "thought" processes. For the

reasons set out below, we first intend to develop these ideas into artificial neural network controller architectures for mobile robots, which will build "ego-centric" representations of the environment and themselves, and which will display properties emulating many of the key functional and phenomenal characteristics of consciousness. We will then progressively de-emphasize the robotic aspects of the implementation, with the aim of arriving at a purely computational system with the desired characteristics of a conscious artificial intelligence. This project is clearly very high-risk – most of this is unknown territory, and the technical demands are formidable. Nevertheless, we believe that the project is feasible at this time, and that the potential rewards are so great that it should be carried out.

Human and Artificial Consciousness:

The search for adequate explanations of human consciousness has recently received much attention (Chalmers 1996, Crick 1994, Dennett 1991, Humphrey 1993, Kelly 1955, McGinn1991, Nagel 1982, Penrose 1989 and 1994, Searle 1992) and indeed our collaborator (Koch) has been at the forefront of such investigations (Crick and Koch 1995, 1998, 1999, and Koch 1994a, 1998a, 1998b). However, the theoretical and practical study of artificial or machine consciousness has received relatively little attention; most of the work has been done by our colleague Aleksander, in the development of his neural state machine "Magnus" and its successors (Aleksander 1993,1994,1996), and by this proposal team's collaboration and interaction (Holland and Goodman 2004).

We endorse the view of Crick and Koch (1998) that it is unprofitable to attempt a formal definition of consciousness at this stage, because of the dangers of premature definition, but that it must involve selective attention, short term memory, and access to the planning stages of the brain. Our approach is constrained and informed by the current state of scientific knowledge in the area. Until the last couple of decades, the field of consciousness was mapped out by philosophers and psychoanalysts, primarily centered around subjective experience, and dominated by argument and speculation. Since then, a tide of neuroscientific data has washed away the traditional content and structure, replacing it with a wealth of often startling findings, such as Sperry's split-brain work (Sperry 1977) but as yet providing no unifying theory.

However, it is still possible to be reasonably certain of a number of facts about consciousness:

- It is a function of neural activity within the brain. (Crick 1994)
- It arises in all normal humans as a function of time, experience, and perhaps interactive human communication.
- It appears to depend on sensory input from the body, since damage to structures interfacing such input with the brain leads to disorders, or the disappearance, of consciousness. (Damasio 1999).
- It exists even when no cognitive processing is taking place, or when some aspects of cognitive processing are impaired by structural or functional damage or abnormality. (Damasio 1999).

- Its contents are normally dominated by the cognitive processing of representations, especially those involving language, logic, and symbolic processing, and by the registration of feelings. (Baars 1988)
- It is closely associated with a process or structure normally referred to as the conscious self, which develops as a function of time and experience, which appears to be stable over extended time periods, and which appears to be the site of cognitive events and the subject of feelings.
- It is possible to make "the self" the subject of conscious attention; this is normally known as self-consciousness.
- The conscious self appears to itself to have a voluntary capacity, which can be used to direct attention in cognitive operations, and to initiate and direct action.
- Because consciousness is a property of highly evolved systems, it must bring some benefit to its possessor.

It is in the light of these observations that we have developed our unique approach. We aim to construct artificial systems, which reflect the characteristics listed above, and which can be used as tools to investigate whether the observed processes and activities can be said to correspond to "what is meant by being conscious".

Approach to Designing a Cognitive Machine:

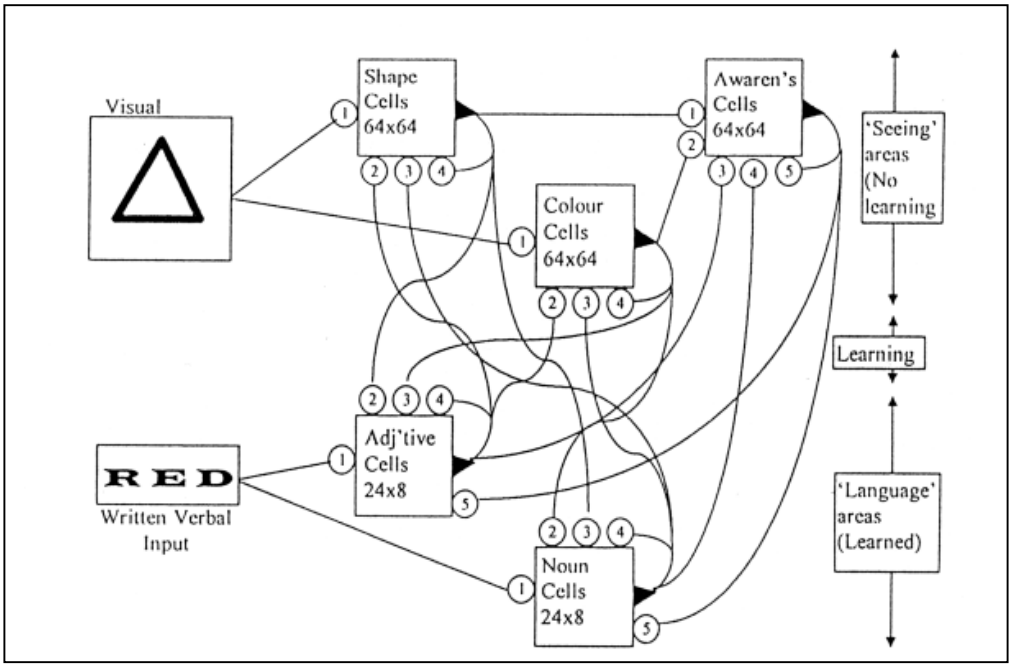
Our approach combines several technologies: Learned neural network controllers, robotics, neuromorphic and other machine sensory systems, visual and other sensory learning, and insights from the biological origins of consciousness. The course we have mapped out is at first sight rather unusual. That is, most of the questions are about information processing, but the answers will be sought initially via robotics. There are several reasons for this, but all ultimately derive from the same fact, that embodiment seems to be crucial for both the development and the maintenance of consciousness, and much of the content of consciousness is related to bodily sensations and actions. In fact, one of the most surprising findings to emerge from consciousness research in neuroscience has been the discovery that consciousness appears to be greatly concerned with, and extremely dependent on, the processing of bodily sensations (Damasio 1999). Such sensations may arise from the state of the body itself, or from the effects on the body of its interaction with the environment, or from changes directly produced in the body by brain activity such as emotion. No-one knows at this stage whether this intimate involvement with bodily sensation, and indeed with embodiment per se, is necessary for the development and support of every possible form of consciousness, but it is certainly so for human consciousness, and this is the only model available for us to follow. At the same time, much of the content of consciousness is to do with the planning of actions in the world – actions that are executed by the body, and that often have (desirable) effects on the body. We therefore propose to study the initial emergence of phenomena analogous to consciousness by building and studying artificial agents which are richly endowed not only with the ability to sense their environment, but also with the ability to sense their own bodies, that is, robots, and in particular robot controllers.

It would of course be possible in principle to develop these concepts using simulated agents, with simulated bodies, operating in simulated environments. Our simple initial experiments have done just this. However, we seek to utilize real robots, with real vision systems, for four reasons. First, much of our ability to deal with the concept of consciousness depends on intuition based on our own experience of consciousness – this is the insight underpinning Crick and Koch's (1998) disinclination to attempt a tight definition of the phenomenon. This intuition exists in the context of embodied systems (ourselves) operating in the real physical world, and we foresee difficulties in applying it to the interpretation of behavior of simulated agents in a simulated and unreal world, beyond the initial research stages. Second, it is clear that a certain richness of internal representations is crucial for the emergence of consciousness, and we can be sure that the real world is rich enough both to enable and to require such representations, whereas a simulated world might not be. Our previous work at the Caltech NSF Engineering Research Center on neuromorphic sensing systems (Braun 2000, Dickson 2000, Gupta 1996, Goodman 1996, Higgins 1994, Keaton 1997, Koch 1994b, Koch 1996, Koch 1999, Kreiman 2000, Lee 1999a, Lee 1999b, Moore 1991, Zeng 1994), and on robotics (Beckers 1994, Holland 1992, Holland 1996, Kyberd 1995, Schoonderwoerd 1997) convinces us that only real robot experiments can capture this richness. Third, there is the simple fact that, given the current state of the art in simulation and in robotics, it is far easier to use real robots. Fourth, and perhaps most importantly, if our proposed robot controller architectures do not "work" – it will be painfully and immediately obvious in a real environment with real robots – they will crash or fail to perform their learned task. Nevertheless, we emphasize that we are not describing a program aimed purely at research into robotics. Far from it: our long-term aim is to develop the technology for embedding artificial consciousness and cognitive processing into computers. Once we have successfully established the structures responsible for a robotic consciousness, we will tackle the challenging problem of migrating the architecture into a purely computational context.

One of the most difficult problems with investigating consciousness in humans (or animals) is that it is essentially private. It is generally agreed that the existence of conscious phenomena in a system in which they are not known to occur cannot reliably be inferred either from action, or from reporting by the system. (This issue has been examined by philosophers of consciousness using the so-called zombie thought experiment, but is also of concern in the areas of clinical neurology and animal rights (Dennett 1991). And while work on the neural correlates of consciousness is advancing, it is still impossible in practice to identify any type of neural activity as being necessarily associated with consciousness. To some extent, this is due to limitations in recording brain activity, but it is also doubtful whether even perfect access to brain activity would enable the identification of such processes, given the current state of knowledge.

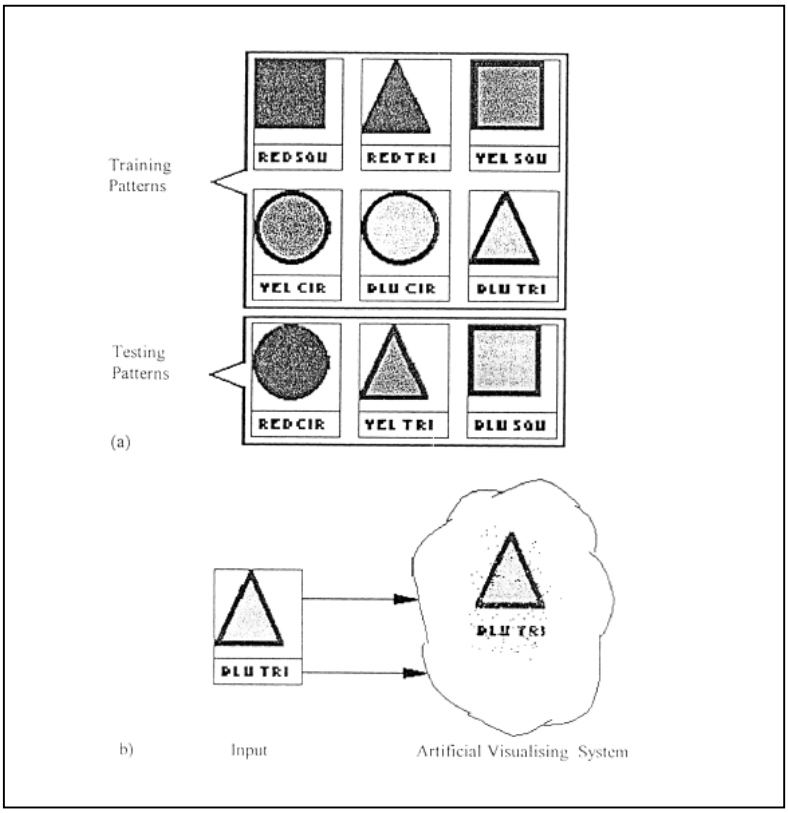
In dealing with possibly conscious activity in artificial systems, the situation may, surprisingly, be rather better. As well as having full access to behavior and reporting, we can also have access to all internal 'brain' activity at any required resolution. Critically, we can also have a full record of all previous behavior, reporting, and brain activity, and we can carry out experimental manipulations that would be technically or ethically impossible with humans or animals. What we require is the ability to identify configurations of brain activity as constituting the systems' *private* "thoughts", and to characterize and represent those thoughts in a *public* way. Searle (1992) has referred to a hypothetical instrument able to do this with humans as a "cerebroscope";

others have called it "the secret policeman's brain scanner"). We propose to deal with this by adapting and extending a solution developed by Aleksander (1999). In an appropriately structured system, exposure to sensory experience builds internal representations of what has been experienced. The problem is this: given a particular sequence of activity in the robot's "brain", how can we know whether it involves such a representation of experience, and how can we know the experience to which it corresponds? Aleksander dealt with this in the realm of simulated visual sensing by constraining his system to develop representations of visual images (iconic representations) which were themselves topographically arranged in the form of images; this enabled the experimenter to recognize the *public* aspects of these representations directly. These representations were also used by the system in a way which exploited their resemblance to images: when the *private* representations were evoked in the absence of direct visual stimulation, in various situations analogous to imagination, they could then be processed or used by the system in the same way as real images. As an example of this architecture and approach we show the computer simulation due Aleksander (1999) which illustrates visual or iconic learning, awareness, and the ability to visualize previously unseen objects in an "awareness" area from their sensed features. The system is "iconically trained", that is, the internal representation in the awareness and other neural areas is an image or "visualization" of the object, and is thus "public". The architecture is shown below.



In this simple example the input to the architecture consists of only two input "sensory" modalities. A *visual* input area in which objects can be displayed, and a *written verbal* input area in which visual text labels can be inserted and displayed. These input areas project into specialist feature detector neural areas. These specialist neural areas have individual architectures, distinct from the overall system architecture, which can implement quite complex modalities such as local feedback, reentrant feedback, and self-adaption. The input objects have *features* in different *categories.* E.g. red in color, square in shape, smooth in texture etc. The specialist visual neural areas have been trained to detect these features, and output the appropriate value. The specialist verbal areas detect such features as "noun" or "adjective". The

specialist neural areas project into a (visual) awareness area, in which the object is visualized. In addition, there is feedback between all the five (non-input) neural areas.



The system is presented with an incomplete training set such as red square, blue circle, etc in its visual and textual input areas, as shown at left. The learning problem is such that when presented with test objects, the system must visualize them correctly in its awareness area. Most importantly, when presented with a yellow triangle at its input, even though it has never seen such an object, it must display a yellow triangle and its verbal label in its visualization area. Note that this is much harder than what is called "concept learning" in the literature. The concept of "yellowness" must have been learned from other objects, and it must be correctly blended with the out-of -context shape word to visualize the previously unseen object. In a sense the architecture has to solve what is commonly called the "binding problem" (Zeki 1993). The system was trained in various strategies of "hard" and "easy" learning. That is, with various degrees of real cues and noise as input. The results indicate that the system is capable of visualizing "unseen" inputs in its awareness area with remarkable robustness as claimed above. A pleasingly anthropomorphic result is that a "visualized" object in the awareness area has more "noise" than a real "seen" object.

We propose to retain both the public and private aspects of this strategy; however, we believe that this particular implementation of the idea imposes limitations on what can be represented successfully. Aleksander's images are static, and it is difficult to see how the technique of iconic representation could be applied to sensory inputs unsuitable for representation in topographical form – for example, sound, or non-visual feedback from motor activation. Therefore we also propose to investigate an alternative strategy of allowing the system to form its own non-iconic *private* representations of both static and dynamic inputs, capable of being exploited by the system as in Aleksander's work, but also training a conventional neural network to produce *public* reconstructions of the inputs to which these representations appear to correspond – a simple process of inversion. (Such an inverse network would be trained by supplying the internal activity as input, and the simultaneous sensory input as the desired output.)

An immediate objection to this program might be that much of our conscious thought consists of abstractions or relationships that do not correspond to any concrete visual or other image, and so the representational capacities of the scheme will prove inadequate to capture anything deserving the name "mental life". The counter-argument to this is illustrated by the autobiography of a remarkable autistic American woman, Temple Grandin, (Grandin 1996) perhaps best known because she was studied by the distinguished neurologist Oliver Sacks, in his paper "An anthropologist on Mars" (Sacks 1996). Grandin entitled her autobiography "Thinking in Pictures" because that is what she does: in spite of the fact that she has mastered spoken and written language, almost her entire mental life consists of pictures – still and moving. Most important, she has described how her use of words expressing abstract relationships, such as "over" and "under", is accompanied by stereotyped visual representations of the concept, such as an image of herself as a young girl under a particular table in a particular room. Although Grandin is certainly very odd in many respects, no-one, not even Sacks, has ever expressed any doubts that she is fully conscious. (She is also highly intelligent and talented: she is a professor of animal science at Colorado State University, and almost half of the cattle-handling equipment in the US uses her designs). Because of Grandin's revelations, we can have some grounds for optimism that any emergent internal activity corresponding to relational concepts may be accompanied by an internal image capable of being decoded by our inversion network.

Learning Robot Controllers:

At the most fundamental level, learning in robots implies learning the robot controller, and moreover that controller is "intelligent" according to some criterion. Some points to bear in mind about robot controllers are:

- Animal and human brains evolved to control behavior in a changeable and partially knowable environment.
- The goal of the controller is to produce the agent's next action.
- The agent uses sensory input, memory, goals, drives, to produce the correct action given the current state of the environment.
- There is only one action at a time.
- Incorrect or multiple actions are very obvious and can damage the robot quickly. (Parkinson's, Huntington's, Tourette's)
- The action may change the environment.
- Good control requires the ability both to predict events, and to exploit those predictions.
- Controllers are layered in increasing levels of abstraction.
- The best such control systems known to engineers are adaptive model-based predictive controllers.

Controllers should be able to:

- Learn models of the environment, the self, and of the interaction of the self with the environment.
- Adapt models automatically based on experience.
- Deal with novel situations automatically, and assimilate the new experience.
- Manipulate models internally to plan actions and goals.

- Make their internal models and reasoning visible in human terms.
- Be able to interact, model, and collaborate on tasks with other similar agents.

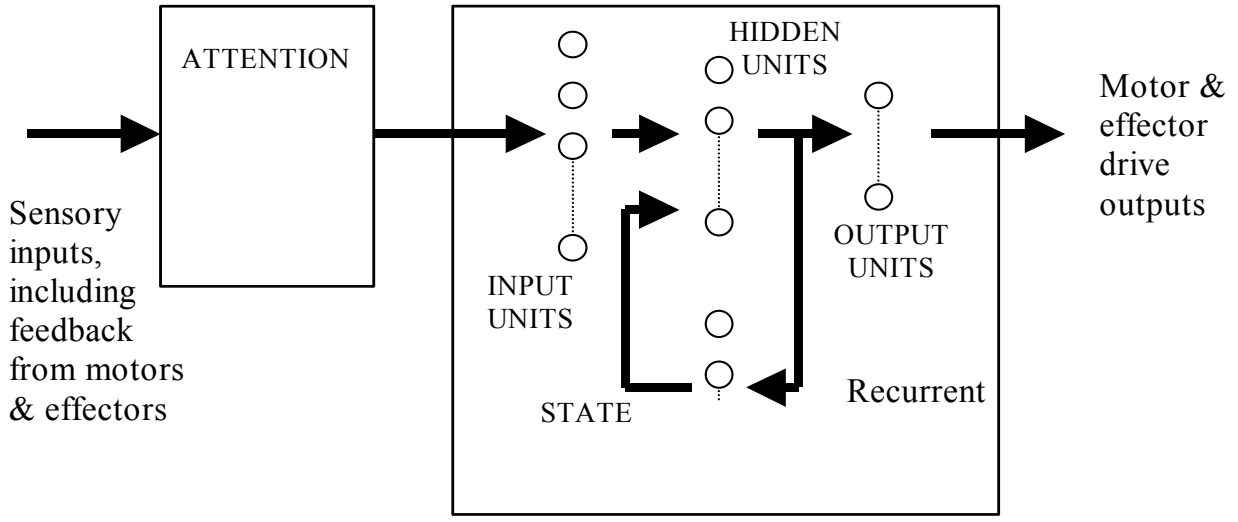
The components needed to implement real time intelligent adaptive controllers are:

- Learning – Adaptation- Reinforcement
- Explicit internal representations
- Environment models - Self models
- Model based predictive control
- Novelty detection
- Attention - Awareness
- Neural Networks - Genetic Algorithms
- Sensory processing, of which vision is the most important

Our approach to learning such robot controllers follows. The controller architecture is a novel neural layered architecture that approaches robot control by layering controllers of increasing abstraction in much the same way as a protocol stack layers functionality in communications systems. The input sensory data is refined and abstracted into higher level concepts by higher level layers. For example, the lowest levels extract "features" from the raw vision input and implement "reactive" control such as crash obstacle avoidance. Higher layers deal with learned feature sequences and "state", and are capable of executing "plans". Information flows both up and down the controller stack, and each layer is capable of taking direct control of the output which at the simplest level is control of the robot's motors. Information flowing "down" the stack allows higher layers to "modulate" the output of the lower layers. In addition, the controller is a neural controller the operation of which is normally "hidden" from us. By implementing inverse modeling, the internal workings of the controller are visible to us and hence credit can be assigned.

A Generic Neural Controller Architecture:

The basic computer architecture for machine consciousness we propose develops from our above hypothesis and approach, and is shown at the system level in the figure below.



The characteristics of this architecture are as follows:

- The controller of the robot is a neural network with recurrent feedback, capable of forming internal representations of sensory information in the form of a neural state machine.
- Sensory inputs (vision, sound, smell, etc) are fed into the controller, including *feedback* signals from the motors and effectors.
- Controller outputs drive the locomotion and manipulators of the robot.
- The neural controller learns to perform a task, using NN and GA techniques.
- Novel inputs that are unrecognized must be adaptively learned by the model.
- The model learns continuously over sequences of actions in time via reinforcement learning, supervised learning, or mimicing a human controller.
- The model continuously refines itself to improve its prediction accuracy.
- But - the internal model of the controller is ***implicit*** and therefore ***hidden*** from us.

We assume that the brain or controller of the robot is an artificial neural network with recurrent feedback, capable of forming internal representations of sensory information in the form of a neural state machine. Sensory inputs (vision, sound, smell, etc) from sensors are fed to this structure via a gating attentional mechanism, that can select inputs to flow through to the controller. (We defer for now the question of how this is controlled). Sensory inputs also include feedback from the motors and effectors, thus giving the robot sensory information on "self". This enables the neural controller to implement a sensory-based ego-centered representation of the environment, and eventually of itself. Output actuation drives the locomotion and manipulators of the robot. The neural controller "learns" to perform its goals, which have been either programmed in, or learned using standard neural network techniques, or "evolved" over "generations" of interaction with the environment via genetic algorithm or evolutionary search techniques (Nolfi 2000).

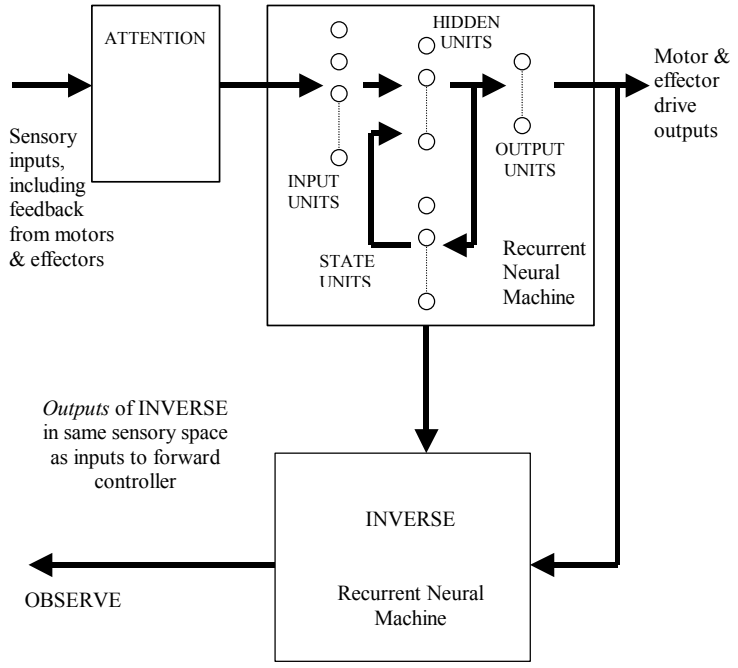
As well as gating the various input channels on and off, the attention module can also shut off all inputs, perhaps apart from a certain amount of noise. (As well as gating off all inputs, the attention module should be able to gate off the outputs, for reasons that will become clear shortly.) The neural state machine is then free to run as a purely state-determined dynamical system. It is well known that the activity in such systems is highly non-random, and tends to converge to attractors related to those formed under the influence of inputs and consolidated by learning. Within the framework of his iconic learning scheme, Aleksander (1996) has shown how the functions of prediction and imagination can be implemented using this scheme; an imagined action is then followed by its predicted environmental consequences, setting the stage for another imagined action, and so forth. This constitutes a potential mechanism for exploring possible sequences of action and evaluating the consequences, and then releasing the motor inhibition to execute an appropriate or desirable sequence in the real world – a classic planning and execution scenario.

There are some interesting consequences of this architecture. In "normal" mode the controller is producing motor signals based on the sensory input it "sees" (including its own motor and effector feedback). Thus it is just like a standard reactive neural controller and can go about its mission according to its learned goals. In "thinking or planning" mode the real world is

disconnected from the controller input, and sequences of planned action towards a goal can take place in "mental space". The actual motor outputs are disconnected from the motors, so that the robot does not actually move; if this were not done, the robot could execute an inappropriate sequence and damage itself or waste energy, and the whole point of thinking rather than acting is to avoid these contingencies. It is interesting to note that there are a number of pathological conditions in humans that correspond to a failure to prevent thoughts or dreams from becoming actions. For example, in Tourette's syndrome, sufferers are unable to refrain from producing utterances and actions which others would inhibit because their consequences are negative. In other conditions, patients and normals may act out components of their dreams while still asleep, often waking in the process. And we are all familiar with sleeping dogs yelping and scrabbling with their paws, presumably while dreaming. (Sleepwalking, and in particular murders committed during sleepwalking, are probably instances of dissociation rather than acting out dreams, as cases are typically characterized by a complete absence of memory for the incident.) The details of a credible neural mechanism for switching this motor engagement on and off appropriately have not yet been worked out in the context of Aleksander-like systems; this is one of the sub-goals of this research program.

Understanding the Controller:

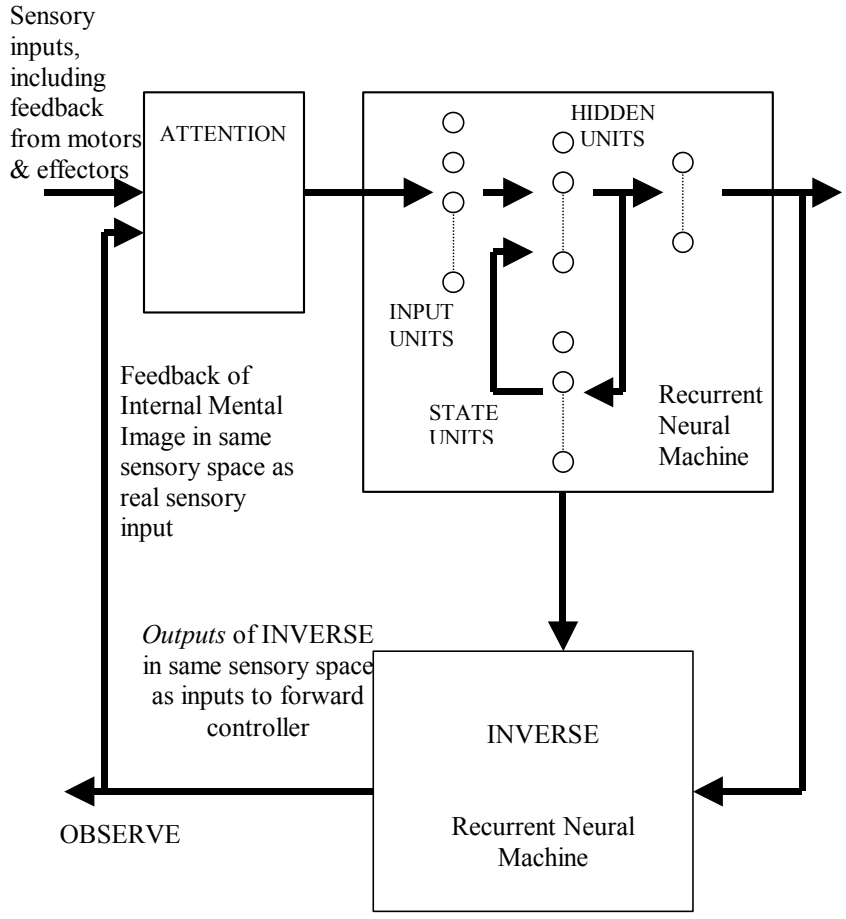
Let us now make a crucial modification of this architecture so that in addition to the usual sensory-motor neural controller, a second recurrent neural network exists, hidden from the first system, which learns the *inverse* relationship between the internal activity of the sensory-motor controller (the hidden and the state units) and the current and previous inputs and outputs. This is illustrated in the diagram below. This mechanism will allow us to represent the hidden state of the sensory-motor controller in terms of the closest corresponding sensory input coming from the real world. Thus we may claim to know "what the robot is thinking", by expressing its "*private* mental image" in terms of the "*public* sensory image" which is visible and available to us. As is standard in learning inverse networks, we require that the controller be learned first, and that, once this is learned and reasonably stable, the inverse can be learned. The advantage of this method is that we are no longer constrained by Aleksander's method of iconic



learning and internal representation, and so we can use any suitable neural model for the sensory-motor controller.

Inverse Predictor Architecture:

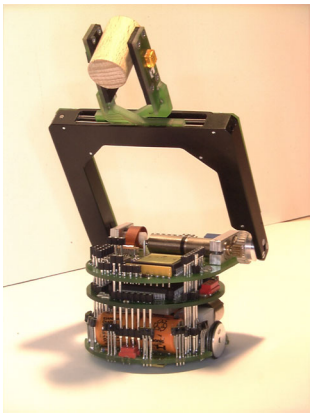
Let us make a further crucial modification: let us incorporate the inverse network into the functional architecture of the system, so that the reconstructed sensory input is not just made available to *us*, but is also available as an *input* to the system via the attention module. This is shown in the diagram below.



There are now some further interesting consequences of this architecture. In "normal" mode, as before, the controller is producing motor signals based on the sensory input it "sees" (including motor/effector feedback). Thus it is just like a standard reactive neural controller and the robot can go about its mission according to its learned goals. The inverse allows for detecting mismatch between a predicted and an actual sensory input (a vital function currently thought to be associated with the cerebellum), or allowing the attentional mechanism to guide the trajectory of the system during planning, just as it would during execution. The inverse system allows us (humans) to "observe" the controllers hidden state, in terms of its sensory input representation, so that we can see what the machine is "thinking" in terms of the "mental images" of the environment. In "thinking or planning" mode the real world is disconnected from the controller input, and the mental images being output by the inverse are input to the controller instead. Thus sequences of planned action towards a goal can take place in mental space, and executed as action. Note that by switching between normal mode and "thinking" mode in some way, we can emulate the robot doing both reactive control and thinking at the same (multiplexed really) time. That is, like humans do when driving a car on "automatic" while "thinking" of something else. In "sleeping" mode we shut off the sensory input and allow noise to be input.

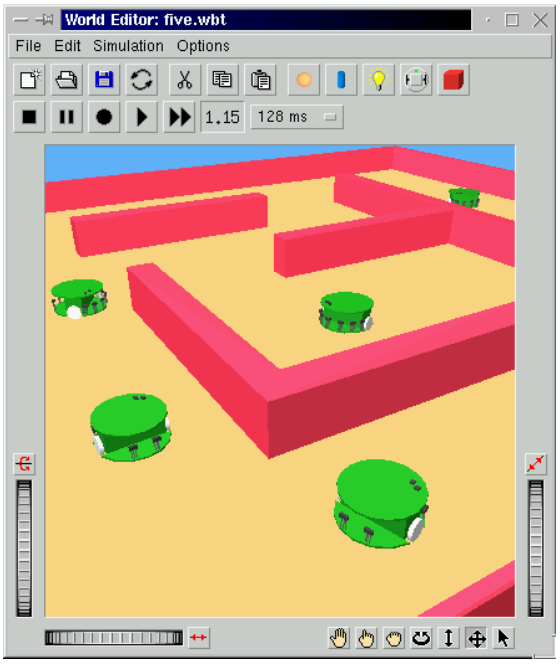
Then the inverse will output "mental images", which themselves can be fed back into the input (because they have the same representation) producing a complex series of "imagined" mental images or "dreams". Note that we can use this "sleeping" mode to actually learn (or at least update) the inverse. The input noise vector is a "sensory input" vector like any other (whether it is structured accordingly or not), thus the inverse should be able to output this vector like any other from the state and motor signals. Thus we can use the error to update the inverse. If we do not disconnect the motors during "dreaming" we will have "sleepwalking". If we assume that the controller is continually learning, then the inverse must be continually updated. If they get too much out of synchronization we could get irrational sequences in "thinking" or worse in execution mode, an analog of "madness".

An Experiment:



To illustrate our approach in making the internal model visible in real robots, we describe an experiment we have performed using an embodied robot simulator "Webots" (Michel1998), and real Khepera robots (Mondada1993). In this experiment, we utilize a controller model based on (Linaker2000) which is much less powerful than the recurrent controllers described above, but allows us to illustrate the principle, and in particular makes "inversion" of the forward controller extremely simple.

The simulated robot is modeled on the Khepera robot (Mondada1993) and features 8 IR sensors which allow it to detect objects, and two independently controlled motors. The picture shows webots in a maze like environment. There are six IR sensors arranged on the forward semicircle of the robot and 2 in rear. The motor drive signals are available. The input feature space of the robot is the 10 dimensional vector of IR sensor values plus the motor drive *signals.*



The crucial simplification we make is that the controller will learn its representation *directly* in the input space. Thus there is no inverse to learn - the internal representation learned by the robot is directly visible as an input space vector.

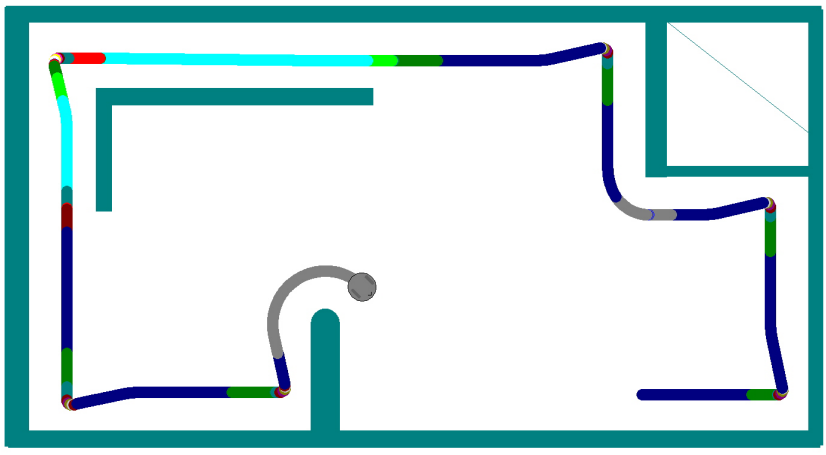
The first phase is to learn or program the forward model or robot controller. Simple behaviors such as collision avoidance or seeking a goal can be programmed in, and more complex behaviors can be developed using behavior based robot controllers (Brooks1986, Mataric1996). In our

experiments controllers are learned or evolved (Banks 2000). In this simple experiment we program in a simple reactive wall-following behavior, rather than learn a complex behavior. The robot starts with no internal model, and adaptively learns its internal representation in an unsupervised manner as it performs its wall following behavior. Learning is accomplished as follows. At every time tick the input vector consisting of the sensor and motor signals is input to the learning controller. The controller maintains an input buffer of N=10 input vectors, and smoothes the current input according to a weighted average of the vectors in the input buffer, in order to cope with noise. The controller has four user defined parameters: a novelty criterion δ, a stability criterion ε, a buffer size *n*, and a learning rate α. In operation the controller network maintains a set M of "concepts", where each concept is represented using a model vector directly in input space. If the input constitutes a novel and stable situation, i.e., the inputs reflect a previously unencountered concept, an additional model vector is allocated and initialized to match the particular input. That is, the set of model vectors M is extended to represent an additional concept, with an accompanying model vector. Additional model vectors are only allocated when novel and stable inputs are encountered, i.e., when the following criteria are fulfilled:

- The input is considered as *novel* if the Euclidean distance between the existing model vectors and the last *n* inputs, compared to the distance between the moving average and the last *n* inputs is larger than the distance δ.
- The input is considered as *stable* if the difference between the actual inputs and the moving average is below the threshold ε.

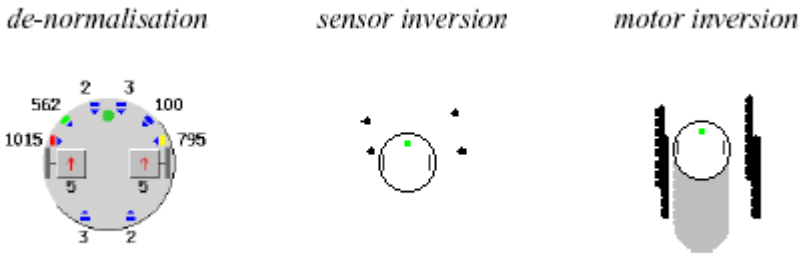
If both the stability and novelty criteria are met, the filtered input is incorporated as an additional model vector. Each time step, a winning model vector is selected, indicating which concept the filtered input currently matches: If the winning model vector matches the filtered input very closely, the filtered input is considered to represent a "typical" instance of the concept, and the model vector is adapted to match the input even closer by an amount dependent on the learning rate. The adaptation is similar to the adaptation of model vectors applied in Learning Vector Quantization (Gray 1998).

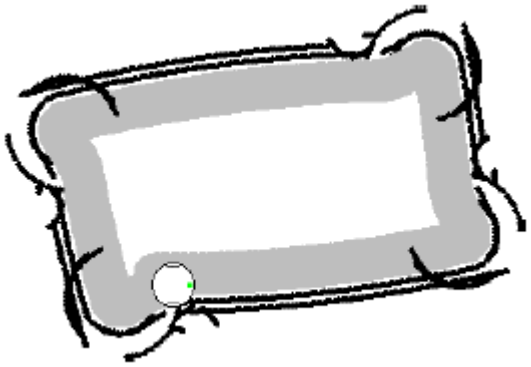
The picture below shows the result of a simulation experiment. The robot has learned a set of concepts (represented by the different colors left in its track). The concepts correspond to "wall on right" (blue), "wall ahead" (green), "right corner" (gray), and "corridor" (light blue). By observing the sequence of winning model vectors (colors) in time we can directly observe what the robot's internal representation is at that time.



The extracted 10-dimensional model vectors can be denormalized to get the actual sensor readings and motor commands for

the concept. By analyzing the sensors, and building up a distance-vs.-activation table, we can find out which distance a sensed obstacle needs to be at, in order to produce a given activation on the sensors as shown below.





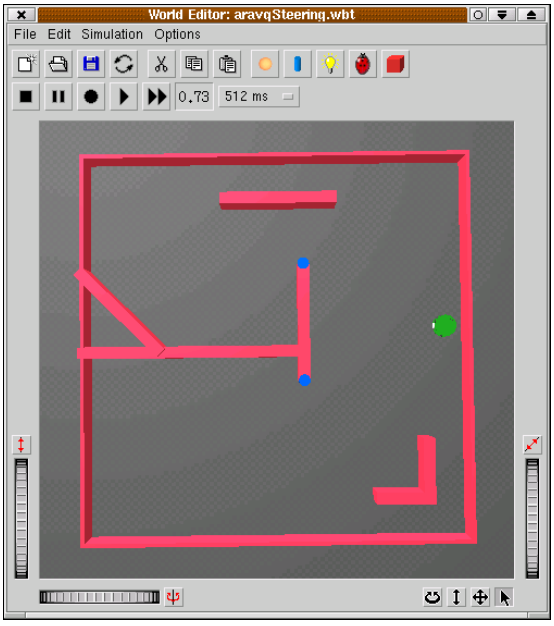
This table can then be inverted in order to find the appropriate distance to the wall when we have a given sensory activation, as is the case with our model vectors. That is, we can plot the location where the robot "thinks" walls "ought" to be, given a certain activation pattern on the sensors. That is, an ego-centric map of the environment can be built up as shown below.

The local representation of detail is surprisingly good; however the global map is distorted. This is primarily because of errors of rotational movement – a well known problem in robotics

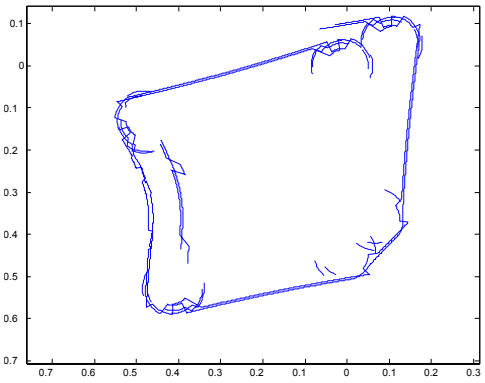
We next demonstrated that the learning algorithm could function in the *real* world by implementing the concept learning and mapping algorithms on a real Khepera robot as shown below. Again, the algorithm proved very robust to real world sensor and motor noise and real world environments.

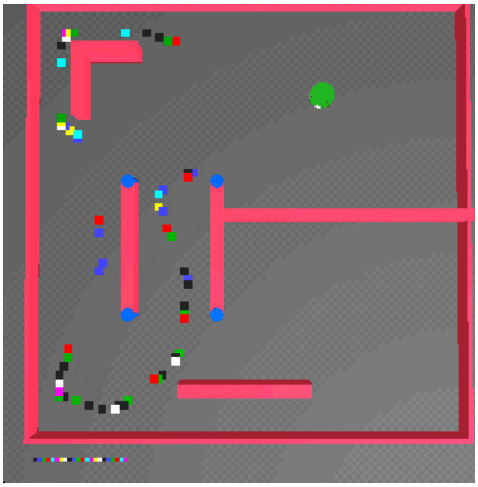
The next crucial step was to allow the learned model, both in simulation and for real, to control the robot. To do this the "teacher" wall follower is turned off, and the robot *sees* "concepts" as it moves. At each time step the closest learned model vector to the sensory feature seen is chosen. The model vector motor drive values are then used to *actually drive the motors* for this time step. This worked surprisingly well. The robot was able to run under "concept" control without any "crashes". Not only did this work for models learned in simulation and on the real robot, but also for models *learned* in the simulation but *run* on the real robot. Furthermore, the environment used for learning was different to that used in execution mode showing that the learned knowledge could be generalized and used in a new previously unseen situation.

For example, the robot learned on the simulated environment shown at left below. The learned model was then downloaded into the real robot shown at right below, and run.



Not only did the controller work in this new real (and previously unseen) environment, but a passable ego centric map of the new environment below was produced. This shows that knowledge can be transferred from one environment and utilized effectively in another completely new environment.



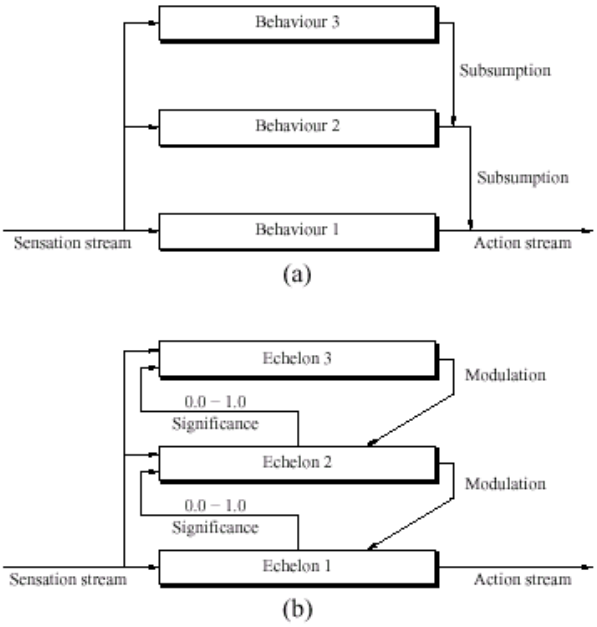


We next implemented an example of manipulating the model mentally in order to allow the simulated robot to make a decision. First we take the sequence of learned model feature vectors and cluster sub –sequences into higher-level concepts (for example: Green-Purple-Blue = Left Corner). Then, at any instant we ask the robot to go to "home". The robot then runs the model forwards mentally to decide if it is shorter (in terms of "concept" sequence length) to go ahead or to go back, and then take the appropriate action. Again this worked robustly. We have thus demonstrated the use of internal models in "planning".

Finally, we implemented a more complex controller and environment. A Braitenberg obstacle avoider was implemented as a teacher, and the environment was as shown above. This resulted in more (22) learned models, which is still very low in terms of complexity of robot behavior.

Higher level controllers:

The controller architecture we propose can be extended by *layering* controllers of increasing abstraction in much the same way as a protocol stack layers functionality in communications systems. The lowest level operates at the ms timescale of sensors and actuator control. The highest levels operate at symbolic levels and much longer goal-driven timescales. The input sensory data is refined and abstracted into higher level concepts by higher level layers. For example, the lowest levels extract "features" from the raw vision input and implement "reactive" control such as crash obstacle avoidance. Higher layers deal with learned feature sequences and "state", and are capable of executing "plans". Information flows both up and down the controller stack, and each layer is capable of taking direct control of the output which at the simplest level is control of the robot's motors. Adjacent layers *modulate* (Linaker 2002) the predictions of higher and lower layers, as opposed to subsumption (Brooks 1990). Information flowing "down" the stack allows higher layers to "modulate" the output of the lower layers. In addition, the controller is a neural controller the operation of which is normally "hidden" from us. By implementing inverse modeling, the internal workings of the controller are visible to us and hence credit can be assigned. This type of controller should be capable of solving much more difficult tasks such as delayed response tasks – e.g. the road sign problem at left, learned using delayed reinforcement learning.



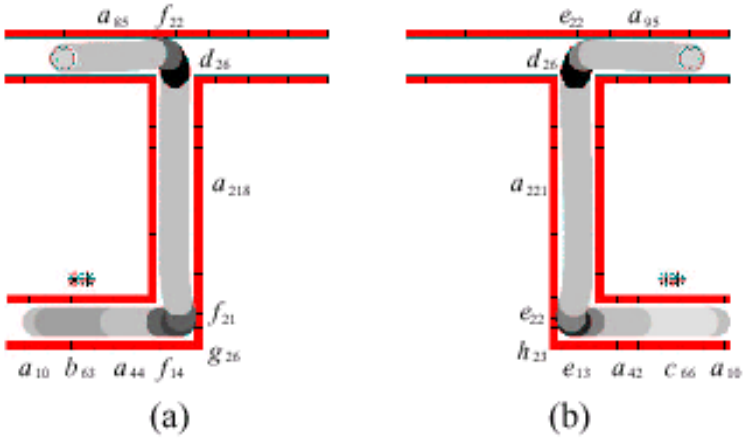


Figure 7: Two examples of distracting events (turns) happening in between the stimulus (light) and the cue (junction).

The proposed program of research:

Our basic strategy is to construct an artifact – a robot – which can sense its environment and its bodily changes with a wide range of sensors, and which can control its behavior using biologically inspired artificial neural structures. The welfare of the robot will be a function of its behavior within the dynamic environment, which will contain other similar robots providing opportunities for cooperation and competition. The environment will be so structured (rich hidden state, requirements for sequenced actions etc.) that good performance will only be achievable if the robot learns to operate at a sufficiently high cognitive level. The robot will be provided with additional internal structures allowing us to monitor its internal processes, especially those in the neural structures, and enabling the representation of the activity within those structures in a form intelligible to us, in visual and auditory displays.

We propose to continue using the Khepera robot(s), and the Webots simulator for our experiments. These robots have a number of advantages. They are professionally built and supported, and have a rich suite of sensors (such as vision systems) and manipulators. In addition, their small size allows for "desktop experiments" of considerable complexity. In this way we will gradually build up the complexity of both the robot and its environment.

We will successively provide various versions of artificial neural structures, and will attempt to demonstrate, by relating the monitored activity to the observed behavior, that the robots are capable of achieving the following aspects of machine consciousness:

1. (a) The development of internal representations of the sensory characteristics of elements in the environment, as a function of sensor-related bodily feedback, ranging from raw sensor images, through feature-based or categorical representations, and perhaps even to symbolic representations. (b) The persistence for some time of such representations when input ceases. (c) The production of such representations in the presence of incomplete or ambiguous stimuli. (d) The reproduction of such representations in the absence of input but in the presence of noise. (e) The use of such representations for functional benefit. We expect all representations at this stage to be indexed by their relationship to the 'point of view' of the robot – so-called ego-centered representations. Items (a) to (d) have already been successfully demonstrated by Aleksander in simulation.
2. (a) The development of internal representations of sequential, temporal, and associative relationships between elements in the environment, with elaborations as in 1.b) – 1(e) above. Items (a) – (d) have already been investigated in simulation using Aleksander's (1999) original MAGNUS architecture, with considerable success.
3. (a) The development of internal representations of the proprioceptive and metabolic inputs from the robot itself, and of their temporal and associative relationships, with elaborations as 1(b) – 1(e) above. Such inputs will include the effects of hard-wired responses associated with mission-related contingencies ("instincts") and may include direct input from neural activity. For example, the internal sensing of low battery voltage could be hard-wired to trigger approach to a recharging station, with perhaps closer approaches to obstacles en route than normally permitted. Low battery voltage might also produce changes in motor actuation (low torque, low speed), reduced motor temperatures, increased battery temperature, reduced

range of sensor operation, and so on. The compound internal representation of all these factors would differ in these respects from the normal representation. We would expect these related representations to form a single complex dynamically changing representation, providing the basis for a "self". By arranging for this representation to form in a particular region, and having various connections from that region to other parts of the control system, the 'self' can have the possibility of exciting or inhibiting activity in particular locations, modalities, structures, or functions, depending on the values of its own internal variables at any time. This corresponds in many ways to enabling the "self" to direct attention, select actions etc., as a function of its own state.

4. (a) The development of internal representations of the various movements possible for the robot, such that the instantiation of these representations initially leads to the execution of the movements. (b) The growth of such representations to include the sensory and/or environmental consequences of the execution of the movement. (c) The growth of such representations to include any effects on internal and metabolic inputs. This stage corresponds to the development of instrumental learning in animals; there are also many analogues in the literature on neural networks in the control of articulated structures.
5. (a) The development of the ability of the 'self' to wholly or partly disconnect sensory input from later, dependent, representations. (b) The activation of those representations by noise and/or other input. This amounts to giving the system autonomous control of the "switch" between internal and external sources of images which is controlled by the experimenter in Aleksander's (1994) early work.
6. (a) The development of the ability of the 'self' to inhibit movement. (b) The subsequent development of representations similar to those in 4 which do not involve the execution of movement, but which involve representations of the sensory and/or environmental consequences which would accompany the execution of the movement. (c) As (b) but involving representations of any internal and metabolic inputs.
7. (a) The integration of all of the above, to produce the entirely internal representation of a situation in which the robot represents the making of a movement which alters both the representations of the sensed environment and the internal and metabolic inputs – in other words, in which the robot imagines making a movement, and imagines the consequences of that movement. (b) The use of such "imagined" actions in delivering benefit, by allowing the 'self' to cease to inhibit movement (i.e. to carry out a particular movement) if the outcome of the 'imagined' movement stimulates the 'self' in certain ways – for example, by stimulating the internal representation of successful recharging. This allows the system to explore the consequences of actions without having to try them out in the current situation, marking the important transition from a Skinnerian to a Popperian machine in Dennett's (1995) terminology. This reflects the contents of consciousness in a conscious cognitive agent solving a problem; these contents would be publicly available to the experimenter as a sequence of images, and could be interpreted as a sequence of thoughts. However, we believe it would not be justified to claim this as consciousness, although it would certainly qualify as a precursor.
8. (a) The development by the robot of an internal representation of its 'self' and its behavior. (b) The use of such a representation to improve performance. If it can be reached, this is a crucial stage. In previous stages, the "self" has no way of taking its own existence into account in dealing with imagined situations; having a representation of itself to manipulate is an important advance. Now the behavior of the "self" includes its responses to internal

representations, whether internally or externally stimulated; these responses include the direction of attention and the release or inhibition of movement, and so it may prove necessary to model not only the "self" but also all of the representations on which the "self" acts, and by which it can be affected. The model "self" would therefore not deal with the "real" representations dealt with by the "self"; this has intriguing parallels with the work of Jackendoff, reinterpreted by Crick and Koch (1999), which hypothesizes that we are not directly aware of thoughts, which are produced by an unconscious homunculus, but are only aware of some derivative sensory aspects of thoughts. This train of thought points us towards the possibility that consciousness is associated not with the "self", but with this internal representation of the "self", which operates not in the world of the "self", but in a derived and secondary representation of that world. This stage would at least qualify as a precursor of self-consciousness, and also should display many of the additional features of consciousness set out in Crick and Koch (1999). The remaining two stages, although vital for an understanding of human-level consciousness and cognition, do not involve the qualitative leaps found in 7 and 8:

9. (a) The development of representations of other similar robots which in some way derive from the system's internal representation of its "self" (Theory of Mind). This may drive an ability for imitation – at many levels – which is critical for the emergence of memes (cultural transmission independent of expected or experienced benefits).
10. (a) The development of language based communication between systems, using signs at first, then arbitrary tokens or symbols, and then introducing syntax. This could be expected to drive the development of internal representations and symbolic thought within each system.

These experiments represent the starting point of this research. Other applications could clearly follow once this architecture is proven to work, and be scalable to larger, more complex, and more unstructured environments, with multiple interacting agents.

References

Aleksander 1993 Aleksander, I. and Morton, H.B., *Neurons and Symbols: the Stuff that Mind is Made of.* (Chapman and Hall, London, 1993).

Aleksander 1994 Aleksander I., "Towards a Neural Model of Consciousness", Proc ICANN 94, Springer (1994).

Aleksander 1996 Aleksander I., *Impossible Minds: My neurons, my Consciousness.* (Imperial College Press, 1996)

Aleksander 1999. Aleksander, I., Dunmall, B., DelFrate, V., "Neurocomputational Models of Visualization: A Preliminary Report", In: Mira and Sanchez (eds) *Foundations and Tools for Neural Modeling*. Springer 1999, pp798-805.

Baars 1998 Baars, B., *A Cognitive Theory of Consciousness*, Cambridge University Press. (1988)

Banks 2000 Banks, L., "Internal Representations in Autonomous Robots: A Novel Approach Using A Recurrent Neural Network Architecture", MURF final report, MicroSystems Lab, Summer 2000.

Beckers 1994 R. Beckers, O. Holland, & J-L. Deneubourg, "From local actions to global tasks: stigmergy in collective robotics", Artificial Life 4, eds. R Brooks and P Maes, MIT Press 1994

Braun 2000 Braun J, Koch C and Davis J (eds) (2000) Visual Attention and Neural Circuits. MIT Press.

Brooks1986 Brooks R., "A Robust Layered Control System for a Mobile Robot". IEEE J. of Robotics and Automation, 1986, Vol. RA-2, No. 1, pp. 14-23.

Caprari 1998 Caprari G., Balmer P., Piguet R., Siegwart R., *The Autonomous MicroRobot "Alice": a platform for scientific and commercial applications*, MHS'98, 9th Int. Symp. on Micromechatronics and Human Science, Nagoya, Japan, November 25-28, 1998, p. 231-5.

Chalmers 1996 Chalmers, D., "Facing up to the Problem of Consciousness", in *Toward a Science of Consciousness,* eds. Hameroff, Kaszniak, and Scott (MIT Press, 1996).

Crick 1994 Crick, F., *The Astonishing Hypothesis*, Scribners, New York, 1994.

Crick 1995 Crick, F. and Koch, C. "Are we aware of neural activity in primary visual cortex?", *Nature*, 375: 121-123, 1995

Crick 1998 Crick, F. and Koch, C. "Consciousness and Neuroscience", *Cerebral Cortex* 8: 97-107, 1998.

Crick 1999 Crick F. and Koch C. "The Unconscious Homunculus", in *The Neuronal Correlates of Consciousness* ed. T. Metzinger, MIT Press, 1999.

Damasio 1999 Damasio A. The Feeling of What Happens: Body and Emotion in the Making of Consciousness, Harcourt Brace, 1999.

Dennett 1991 Dennett, D. C. Consciousness Explained (Allan Lane/Penguin, London, 1991).

Dennett 1995 Dennett D. (1995) Darwin's Dangerous Idea: Evolution and the Meanings of Life. Simon Schuster.

Dickson 2000 Dickson, J. A. and Goodman, R. M., "Integrated chemical sensors based on carbon black and polymer films using a standard CMOS process and post-processing," ISCAS2000, 2000 IEEE International Symposium on Circuits and Systems, Emerging technologies for the 21st century Geneva, Switzerland, May 28 - May 31, 2000.

Goodman 2003 Goodman, R.M. "Adaptive Real-time Learning of Robot Controllers," DARPA Workshop on Navigation, Locomotion, and Articulation, November 11, 2003, Arlington, VA.

Goodman 1996 R.M. Goodman & G. Erten, "Analog VLSI Implementation for Stereo Correspondence Between 2-D Images," IEEE Transactions on Neural Networks, Vol. 7, No. 2, pp 266-277, March 1996.

Grandin 1996 T. Grandin, Thinking in Pictures, Vintage, 1996.

Gray 1998 Gray, R.M. and Neuho , D.L, Vector Quantization, In IEEE Transactions on Information Theory, 44(6), 1998.

H. Greenspan, J. Goldberger and L.Ridel. "A Continuous Probabilistic Framework for Image Matching", Journal of Computer Vision and Image Understanding. 84:384-406, 2001.

H. Greenspan, J. Goldberger, A. Mayer, "Probabilistic Space-Time Video Modeling via Piecewise GMM", Accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence (2003).

H. Greenspan and R.M. Goodman, R. Chillappa, C. Anderson, "Learning Texture Discrimination Rules in a Multiresolution System," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 9, pp. 894-901, September 1994.

F. Linaker and L. Niklasson, "Sensory-flow Segmentation Using a Resource Allocating Vector Quantizer, Advances in Pattern Recognition, Springer, pp.853-862, 2000

Gupta 1996 B. Gupta, R. Goodman, F. Jiang, Y-C. Tai, S. Tung, C-M, Ho, "Analog VLSI System for Active Drag Reduction," IEEE Micro, Vol. 16, No. 5, October 1996, pp. 53-59.

Higgins 1994 C.M. Higgins and R.M. Goodman, "Fuzzy Rule-Based Networks for Control," IEEE Transactions on Fuzzy Systems, Vol 2, No. 1, pp 82-88, Feb. 1994.

Holland 1992 O. Holland & M. Snaith, "The neural control of locomotion in a quadrupedal robot", IEE Proceedings-F: Radar and Signal Processing, December 1992.

Holland 1996 O. Holland & C. Melhuish. "Some adaptive movements of animats with single symmetrical sensors", 4th Conference on Simulation of Adaptive Behavior, Cape Cod, 1996.

O. Holland and R. Goodman, "Robots with Internal Models," Journal of Consciousness Studies, Volume 10, No.4-5(2003).

Humphrey 1993 Humphrey, N., *A History of Mind* (Vintage, London, 1993).

Keaton 1997 Keaton, P., Greenspan, H. and R. Goodman, "Keyword Spotting for Cursive Document Retrieval", in Proceedings of the IEEE Workshop on Document Image Analysis (DIA'97), in conjunction with the Conference on Computer Vision and Pattern Recognition (CVPR'97), San Juan, Puerto Rico, 1997, page(s): 74-81.

Kelly 1955 Kelly, G. A., *The Psychology of Personal Constructs* (Norton, New York, 1955)

Koch 1994a *Large-Scale Neuronal Theories of the Brain*, C. Koch and J. Davis, eds., MIT Press, 1994.

Koch 1994b Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits. Koch, C. and Li, H., eds., IEEE Computer Science Press 1994.

Koch 1996 Koch, C. and Mathur, B. Neuromorphic Vision Chips. IEEE Spectrum, 33 (5): 38-46, 1996.

Koch 1998a Koch. C., *Biophysics of Computation: Information Processing in Single Neurons,* Oxford University Press, 1998.

Koch 1998b Koch, C. and Segev, I., eds., *Methods in Neural Modeling: From Ions to Networks*. Completely revised second edition. MIT Press, 1998.

Koch 1999 Koch C & Laurent , "Complexity and the nervous syste",. Science 284: 96-98.

Kreiman 2000 Kreiman G, Koch C & Fried I, "Imagery neurons in the human brain", Nature 408: 357-361, 2000.

Kyberg 1995 P. Kyberd, O. Holland, P. Chappell, S. Smith, R. Tregidgo, P. Bagwell, & M. Snaith., "MARCUS: A Two Degree of Freedom Hand Prosthesis with hierarchical grip control", IEEE Trans on Rehabilitation Engineering, Vol 3, no 1, March 1995

Lee 1999a Lee, D., Itti, L., Koch, C. and Braun, J. Attention activates winnter-take-all competition amongst visual filters. Nature Neuroscience 2: 375-381, 1999.

Lee 1998b Lee DK, Koch C & Braun J (1999) Attentional capacity is undifferentiated: concurrent discrimination of form, color, and motion. Perception & Psychophysics 61, 1241-1255.

Linaker 2000 Linåker, F. and Niklasson, L., Extraction and Inversion of Abstract Sensory Flow Representations. In Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior: From Animals to Animats 6 (SAB2000), MIT Press, pp. 199-208, 2000.

Martinoli1999a Martinoli A., "Swarm Intelligence in Autonomous Collective Robotics: From Tools to the Analysis and Synthesis of Distributed Control Strategies". PhD Thesis, EPFL, Lausanne, Switzerland, 1999.

Mataric1996 Mataric M. J. and Cliff D., "Challenges in Evolving Controllers for Physical Robots". Robotics and Autonomous Systems, Special issue on Evolutionary Robotics, 1996, Vol. 19, No. 1, pp. 67-83.

McGinn 1991 McGinn, C. *The Problem of Consciousness* (Basil Blackwell, Oxford, 1991).

Mondada 1993 Mondada F., Franzi E., and Ienne P., "Mobile Robot Miniaturization: A Tool for Investigation in Control Algorithms". In Yoshikawa T. and Miyazaki F., editors, Proc. of the Third Int. Symp. On Experimental Robotics, Kyoto, Japan, October, 1993, Lecture Notes in Control and Information Sciences, Springer Verlag, pp. 501-503.

Moore 1991 A. Moore, J. Allman and R. Goodman, "A Real-time Neural System for Color Constancy," IEEE Transactions on Neural Networks, Vol. 2, No. 2, pp 237-247, March 1991

Nagel 1982 Nagel, T *The View from Nowhere* (Oxford University Press, Oxford, 1982)

Nolfi 2000 Nolfi, S., and Floreano, D., *Evolutionary Robotics*, MIT press, 2000.

Michel 1998 Michel O., "Webots: Symbiosis Between Virtual and Real Mobile Robots". In Heuding J.-C., editor, Proc. of the First Int. Conf. On Virtual Worlds, Paris, France, July, 1998, Springer Verlag, pp. 254-263.

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

Penrose 1994 Penrose, R. *The Shadows of the Mind* (Oxford U. Press, 1994)

Sacks 1996 Sacks, O., *An Anthropologist on Mars*, Vintage, 1996.

Searle 1992 Searle, J.R. *The rediscovery of the mind* (MIT Press, Boston,1992).

Schoonderwoerd 1997 R. Schoonderwoerd, O. Holland, J. Bruten, & L. Rothkrantz.. "Ants for load balancing in telecommunications networks", Adaptive Behavior, 5, 2, 1997.

Sperry 1977 Sperry, R.W. "Forebrain Commissurotomy and Conscious Awareness", The Journal of Medicine and Philosophy, 2, pp.101-126, 1977.

Winfield 2000 A.F.T.Winfield and O.E.Holland, 'The application of wireless local area network technology to the control of mobile robots', Microprocessors and Microsystems, 23, 10, 2000, pp 597-607

Zeki 1993 Zeki, S., *A Vision of The Brain*, Blackwell Science, 1993.

Zeng 1994 Z. Zeng, R.M. Goodman P. Smyth, "Discrete Recurrent Neural Networks for Grammatical Inference," IEEE Transactions on Neural Networks - Special Issue on Dynamic Recurrent Neural Networks: Theory and Applications, Vol 5, No. 2, pp 320-330, March 1994.

The PI is **Dr Rodney M. Goodman** who is CEO and President of Gaea Corporation. Gaea Corporation provides advanced technology solutions to California High Tech industries, in the areas of Electronics, Communications, Signal Processing, and Intelligent Systems. Dr Goodman is a US citizen. A full CV and publication list is available on www.rodgoodman.ws

Rodney M. Goodman was born in London England, on February 22, 1947. He received his B.Sc. degree with Honors in Electrical Engineering from Leeds University, Yorkshire, UK in 1968, and his Ph.D. degree in Electronics at the University of Kent at Canterbury, UK in 1976, under the supervision of Professor Paddy Farrell. From 1975 to 1985 he was a member of the faculty of the Department of Electrical and Electronic Engineering at the University of Hull, UK. In 1985 Dr. Goodman joined the faculty of the California Institute of Technology (Caltech) where he was full Professor of Electrical Engineering until September 2001. Dr Goodman is currently CEO and President of Gaea Corporation. His clients include several advanced technology start-up companies in the Pasadena area, including Cyrano Sciences, the electronic nose company, of which he is a founder. In addition, Dr. Goodman is a Faculty Associate in Electrical Engineering at Caltech.

Dr. Goodman's current research interests are in intelligent information processing systems, electronic nose technology, distributed communications networks of sensors and actuators, ultra wideband wireless, ad-hoc networks, RFID and RTLS. In addition, novel control architectures for multiple autonomous mobile robots, and machine consciousness are being pursued.

Goodman's research lab at Caltech comprised three groups: the Collective Robotics Group (CORO), the Information Processing Systems Group, and the Neuromorphic VLSI Processing Group. While at Caltech Dr. Goodman developed new error control coding algorithms for VLSI memories which have been implemented on spacecraft missions, and various neural network VLSI implementations, including: neural associative memories with large capacity, artificial MEMS skin chips, and the Caltech silicon nose chip. He also developed new expert system technologies that have been successfully transferred to industry for control and management of communications networks. These include a new class of rule-based neural networks, which feature explicit knowledge in the form of human understandable rules. Robotics activities were focused towards swarm intelligence and collective robotics. Dr. Goodman was the founding PI and director of the National Science Foundation's Center for Neuromorphic Systems Engineering at Caltech. These are national centers of excellence, and the Caltech Center for Neuromorphic Systems Engineering, together with its educational component, the Computation and Neural Systems graduate degree program, is at the forefront of this field. The mission of this center is to develop the technologies necessary for endowing the machines of the future with the human-like senses of vision, audition, touch, and smell and taste (chemical sensing), and applying these to autonomous robots.

Dr. Goodman has consulted for a variety of government and commercial organizations in both the US and the UK, and has a current US Secret Clearance. He is a founder of four advanced technology research and development companies in both the US and the UK, and is currently a consultant to several high technology companies in the Pasadena area including IdeaLab!, Calhoun Vision, and Cyrano Sciences. Dr. Goodman is a Fellow of the IEE, a Chartered Electrical Engineer, and a Senior Member of the IEEE. His honors and awards include two

NATO Senior Scientist Awards and a Research Fellowship of the Royal Society. Dr Goodman has served as North American editor of Neural Computing and Applications, and has served as a reviewer for various IEEE (U.S.A.), and IEE and IERE (U.K.) journals including: IEEE Transactions on Information Theory, Computers, Neural Networks, Pattern Analysis and Machine Intelligence, Proceedings of the IEEE, Proceedings of the IEE, Electronics Letters, and Neural Computation. Dr Goodman has served on various organizing and program committees for: IEEE International Information Theory Symposium, Neural Information Processing Systems (NIPS), NIPS Foundation, International Joint Conference on Neural Networks (IJCNN), Neural Networks for Computing/Machines that Learn (Snowbird), IFIP International Symposium on Integrated Network Management (ISINM), International Symposium of Circuits and Systems (ISCAS), International Workshop on Applications in Neural Networks in Telecommunications (IWANNT), Frontiers in Distributed Information Systems (FDIS), and the International Workshop on "Can a Machine be Conscious" – Cold Spring Harbor. Dr. Goodman has published over 150 technical papers and patents in his areas of expertise.

