Robots With Internal Models A Route to Machine Consciousness?

Article *in* Journal of Consciousness Studies · January 2003

Owen Holland
University of Sussex

Reprinted from:

Journal of Consciousness Studies

controversies in science & the humanities

Vol. 10, No. 4-5, April/May 2003

SPECIAL ISSUE: MACHINE CONSCIOUSNESS

Edited by Owen Holland

Owen Holland and Rod Goodman

Robots With Internal Models

A Route to Machine Consciousness?

New organs of perception come into being as a result of necessity - therefore,
0 man, increase your necessity so that you may increase your perception.

Jalaluddin Rumi, Persian poet (1207-1273)

Introduction

We are engineers, and our view of consciousness is shaped by an engineering ambition: we would like to build a conscious machine. We begin by acknowledging that we may be a little disadvantaged, in that consciousness studies do not form part of the engineering curriculum, and so we may be starting from a position of considerable ignorance as regards the study of consciousness itself. In practice, however, this may not set us back very far; almost a decade ago, Crick wrote: 'Everyone has a rough idea of what is meant by consciousness. It is better to avoid a precise definition of consciousness because of the dangers of premature definition. Until the problem is understood much better, any attempt at a formal definition is likely to be either misleading or overly restrictive, or both' (Crick, 1994). This seems to be as true now as it was then, although the identification of different aspects of consciousness (P-consciousness, A-consciousness, self consciousness, and monitoring consciousness) by Block (1995) has certainly brought a degree of clarification. On the other hand, there is little doubt that consciousness does seem to be something to do with the operation of a sophisticated control system (the human brain), and we can claim more familiarity with control systems than can most philosophers, so perhaps we can make up some ground there.

The starting point we have chosen is at the conjunction of four fairly uncontroversial observations:

- **consciousness is known to arise from the operation of the human brain**
- **of all brains, the human brain has the highest capacity for intelligence**
- **the human brain evolved from simpler brains**
- **the human brain is a control system**

Taken together, these suggest a strategy: step by step, develop a series of control systems capable of demonstrating increasingly high intelligence, and monitor their performance at each stage to see when and if they show any signs of consciousness. The immediate appeal is that it does not involve sitting down and 'inventing' machine consciousness, or human-level intelligence, either of which would be extremely problematical; the technical aspect of this approach involves nothing more than the successive development of a series of control systems for producing increasingly intelligent behaviour, and that is a reasonable goal for engineers. Of course, it may also be completely misguided, in that high intelligence may not automatically entail consciousness, but, given the present state of ignorance about so much to do with consciousness, it seems a risk worth taking.

This paper explores some of the issues relevant to the consciousness-via-incremental-intelligence programme, and makes two key design decisions: to embody intelligence in a physical robot, and to concentrate on the exploitation of internal models. It also presents some speculations about how the pursuit of intelligence in this way may lead us to a system with at least some of the characteristics of consciousness.

Choosing an Architectural Stance: Internal Modelling

The scheme as outlined so far is too broad to define a distinctive and manageable approach. There are a great many types both of control systems, and of information processing architectures; we cannot explore them all, but the choice of one particular type should be principled rather than arbitrary. It should of course hold out the prospect of offering a route to high intelligence, but ideally it should also combine some credible connection with consciousness along with some possibility of being implemented in natural nervous systems. As it happens, there is a very strong candidate principle: the development and use of internal models. There are of course a great many structures in information processing architectures that could be considered as being internal models of one sort or another. In addition, the number of different characteristics of the world, the robot, and their interaction that might be modelled - the physical, abstract, logical, temporal, and other aspects - is effectively unlimited. All we can do in the space available here is to

sketch out some of the kinds of internal models that we believe might be useful within the context of this project, and in particular to identify their connections with control systems, intelligence, the brain, and consciousness. (It is difficult to find a good general theoretical treatment of this field, although it was a major topic of study within cybernetics; Arbib' s *The Metaphorical Brain* [Arbib, 1972] is still more useful than many later texts.) The sense in which the word 'model' is used in engineering is slightly different from its use in everyday speech. The basic characteristic of any model, internal or external, might seem to be that it is some sort of process or structure that in some way resembles whatever it purports to model. However, when the model is to be used for some purpose by the control system of a robot, the model will be represented within some internal information processing system, and the sole requirement is that information processing operations involving the model should yield appropriate outputs in relation to the aspects of the situation being modelled (Minsky, 1968). There is no intrinsic requirement for the model itself to correspond to reality in any other way. (Of course, this does not preclude the use of models which do have a clear resemblance to whatever is being modelled.) For example, in the field of artificial neural networks, a network that has been trained to produce a particular set of outputs in response to a given set of inputs is often said to have learned 'an internal model' of the problem; this internal model is simply a pattern of synaptic weights that happens to give the correct outputs, and it is only in rare cases or in certain special types of networks that the characteristics of the internal model can be related explicitly to the characteristics of the problem. Engineers are often particularly interested in what might be called 'working models', in which induced changes in one part of the model cause all the linked parts of the model to change appropriately and in step with one another, so that the dynamics of the whole system is captured. This is different from having a database of facts, or a look-up table, where a suitably encoded enquiry produces the correct answer without the mediation of any active modelling process. A potential problem with this approach is that the attempt to achieve intelligence in artificial systems by the formation and exploitation of internal models of the world has been going on in one form or another since the Dartmouth Conference of 1955; in fact, it is the programme of what is often referred to pejoratively as GOFAI (Good Old-Fashioned Artificial Intelligence). Its critics, notably Brooks (1991) and Dreyfus (1992), have claimed that the idea of controlling behaviour by building and reasoning over symbolic internal models (or representations) in this tradition is fundamentally flawed. We agree, but point out that other forms of internal models and other methods of exploitation are largely unaffected by their criticisms, and that the success of modelling approaches within engineering shows that complex systems can be controlled in real time by appropriate variants of these methods, as described below.

The Use of Internal Models in Control Systems

In engineering, the system that needs to be controlled - such as an aeroplane, oil refinery, or air conditioning unit - is known as the plant. The control system sends control inputs to the plant that affect the operation of the plant in various ways, and it receives outputs from the plant that it may use, along with environmental information, when calculating the control inputs. The purpose of the control system is to force the plant outputs to achieve or maintain some state, or to follow some desired trajectory. Environmental factors may also produce disturbances, creating the need for control actions. There are two broad classes of control: feedback, and feedforward.

In feedback control, also known as closed loop control, the control inputs are constantly modified as a function of the plant outputs. The simplest and best known form is negative feedback: the control task is to maintain some value of the plant output at a fixed level, and the controller generates a control input that that is a function of the difference, or error, between the measured value and the desired value, and tends to reduce that difference. If there is no error, the controller produces no output. Problems can arise when there are long time delays in the loop - for example, if the sensor measuring the controlled output value is very slow, or if the plant itself takes a long time to respond to a control input. Since the controller is driven by errors, it will not be able to achieve very rapid responses, and other techniques will have to be used.

Modern feedback controllers tend to use models of some kind to improve their performance. Rather than simply measuring the plant output that is to be controlled, measurements of many other variables can be taken so that the current state of the plant can be estimated accurately; this amounts to constructing a kind of model. In situations where the behaviour of the plant is incompletely known, or changes with time, a computer system running a more advanced controller (a so-called adaptive controller) may develop or adjust an internal model of the plant (a process known as system identification); alternatively, the controller may adjust itself so that the effect it has on the plant matches the computed effect on an idealised internal model of the plant. Adaptation is always a much slower process than the control process itself.

In feedforward control, sometimes called open loop control, the control inputs are calculated from the following factors: the desired control value; measurements of environmental variables (e.g. disturbances); and the current state of the plant. They are then passed to the plant, and are not modified by any plant outputs occurring after that time. Because these controllers do not have to wait for the sensed response of the system, they can operate extremely fast. A good feedforward controller will produce a control input that exactly cancels the effects of any environmental disturbance and achieves the target value for the controlled output. For example, a robot arm might have to be moved very quickly to a particular end point; a feedforward controller would calculate from the initial position and the desired position of the arm the exact timing and amount of electric current to be supplied to each joint motor to produce the required movement. This calculation can be thought of as requiring some sort of model of the arm's characteristics - an implicit model, expressed only in terms of the current patterns necessary to produce particular movements. This type of model is known as an inverse model, and is the opposite of a forward model, which yields the movements or sensory conditions that will result from the application of a particular configuration of joint motor currents.

More complex controllers, sometimes known as optimal controllers, or model-based multi-step predictive controllers, do not attempt to keep the instantaneous values of various plant outputs close to some desired profile, but instead operate to minimise the cumulative value over time of some cost measurement associated with the plant. They do this by 'planning' the sequence of control actions stretching into the future that is calculated to produce the lowest-cost sequence of plant changes; they then execute the first action, and calculate the optimal sequence afresh. In order to do this, they need a good model of the behaviour of the plant; again, there are techniques for acquiring or modifying such a model if it is partially unknown or if it changes over time. The most sophisticated controllers currently in use are usually called adaptive model-based predictive controllers.

The controllers mentioned above are usually implemented using conventional mathematical formulations based on linear systems, or on linear approximations to non-linear systems. The theory behind such methods is well developed, and makes it possible to design control systems with proven characteristics of stability and convergence. In the last fifteen years, some less conventional methods have been developed to cope with systems which are highly non-linear, or in which the nature of the system model is unknown. Some types of neural networks are able to learn accurate models of unknown non-linear functions, and have given rise to the field of neurocontrol (Werbos, 1990; 1992). Where a non-linear model is partly known, control systems using fuzzy logic offer rapid and accurate control (Passino and Yurkovich, 1998). The learning abilities of neural networks are sometimes combined with the tractability of fuzzy systems in neuro-fuzzy control systems (Nauck *et al.*,1997). These new methods are highly adaptable, and can be used as standalone systems, or as components of model-based predictive controllers or other techniques originally developed for linear systems.

It can therefore be seen that models of what is being controlled, and how it responds to control inputs, are widely used in control systems; these models are of various types, and are used for various purposes.

What Could a Robot or an Animal Use an Internal Model For?

We have identified four major applications which seem to be relevant for animals and robots, and which align well with our programme: processing novel or incomplete data; detecting anomalies; enabling and improving control; and informing decisions. As there is not the space to deal with them in any detail, we restrict ourselves to giving an example of each.

Processing novel or incomplete data. Any scheme that, in effect, allows a novel input to be mapped to a category, and treated as an exemplar of the category, can be regarded as involving a kind of model. In many animals, exposure to a stimulus with certain fixed characteristics (a 'sign stimulus') may release a stereotyped response; for example, the presentation of a red patch will cause a male robin to attack it, regardless of whether it is another robin's breast (the natural stimulus), a piece of cardboard, or even a Post Office van seen through a window. Regardless of the actual mechanism used, the robin (or other organism or robot) could be said to have some kind of implicit internal model of 'that-which-is-to-be-attacked'. Of course, if the mechanism involved the explicit comparison of a processed sensory input with some template, the template would correspond more closely to the everyday meaning of 'model' .

Detecting anomalies. If a model generates an expectation or prediction about what goes with what, or what follows what, then any unexpected conjunction or succession can be detected by the system of which the model is a part. This has survival value because novelty often indicates danger, or some other situation that is worth attending or responding to. It is in some ways the counterpart ofthe first use of models because it can be regarded as the category of unmodelled occurrences. For example, chimpanzees will tolerate a doll that looks like a normal chimp, but will respond with fear or aggression to a deformed or incomplete doll.

Enabling or improving control. Any of the examples given above in the context of model-based control will suffice.

Informing decisions. A model can be used in many ways to predict the consequences or utility of various possible future actions, and so can be used to guide the generation or selection of such actions. There are two basic scenarios for prediction: single step and multi-step. In a single step prediction, all that is predicted is the consequence of the next action; in a multi-step prediction, what is predicted is the consequence of a sequence of actions. However, a multi-step prediction does not necessarily imply the commitment to an entire sequence of actions, because it can also be used merely to select the first action from the preferred sequence, with the whole simulation and selection process being run after every executed action. This is typical of the approach used in engineering, and in many planning systems in AI. From a cognitive viewpoint, the outcome of any of these processes can be regarded as representing a decision; however, it should be remembered that the selection of one of a number of alternative possibilities can also be achieved by what some might see as a non-cognitive mechanism, such as a winner-take-all network. The role of the model is not to make the decision, but rather to provide information to the decision-making process.

The use of models to inform decisions is particularly relevant to our project, and most of what we want to convey is captured by Dennett in his description of a hypothetical creature, the last of three in an evolutionary sequence (Dennett, 1995). The first, the Darwinian creature, is the basic model. Its responses to its environment are specified by its genes; those examples with genes producing bad responses die, and those with genes for good responses survive to breed, eventually producing a population with better responses. The second, the Skinnerian creature, is capable of learning, and as a result becomes capable of producing better responses if it is not killed by an early bad response. The one we are interested in is the third, the Popperian creature, which is able to preselect its responses so that those likely to kill it are inhibited.

> But how is this preselection in Popperian agents to be done? Where is the feedback (about the quality of the proposed action) to come from? It must come from a sort of inner environment - an inner something-or-other that is structured in such a way that the surrogate actions it favours are more often than not the very actions the real world would also bless, if they were performed. In short, the inner environment, whatever it is, must contain lots of *information* about the outer environment and its regularities. . . we must be very careful not to think of this inner environment as simply a replica of the outer world, with all its physical contingencies reproduced. . . . The information about the world has to be there, but it also has to be structured in such a way that there is a nonmiraculous explanation of how it got there, how it is maintained, and how it actually achieves the preselective effects that are its *raison d'etre* (Dennett, 1995, pp. 375-6)

A more technical theoretical treatment of some of the ways in which techniques derived from control systems might be applied to cognition can be found in Grush (1997; 2002). Clark and Grush (1999) extend the analysis to the problem of achieving meaningful cognition in robots. Although the approach is primarily philosophical, the content could easily be adapted for implementation on a robot, and in that case would probably be very much in line with what is being proposed here.

The Presence and Use of Models in the Brain

There is a considerable amount of evidence that many of the control systems in the brain use models in much the same way as control systems designed by engineers. However, the brain also appears to exploit models in ways that go beyond the current capabilities of engineering systems. This brief review can do no more than scratch the surface ofthis topic, and the examples given below are a sample rather than a balanced view of an extensive literature; see Wolpert and Ghahramani (2000) for a more comprehensive presentation.

A major problem the brain faces it that its sensory systems are relatively slow - visual feedback takes around lOOms - and so feedback control cannot be used to control rapid movements. Since the brain does manage to produce appropriately timed and accurately modulated muscle activations to control rapid movements, it must be using a feedforward scheme with inverse models to do this. The variability of tasks, situations, and bodies means that these inverse models cannot all be innate; plausible schemes have been advanced concerning the ways in which they might be learned.

One characteristic of multi-jointed systems such as the human body is that there are an infinite number of ways of moving a limb to a specified position; the control problem is said to be underdetermined. However, there is a remarkable degree of stereotyping when an individual is asked to repeat a movement, and when different people are asked to make the same movement. This seems to be due to the brain using a form of optimal control, which selects from all the possible trajectories the one that minimises some cost function, and does so by using some kind of model. A lot of effort has gone into discovering the cost function that the brain uses; whatever it is, it tends to produce smooth movements with smooth changes of torque at the joints.

Brains also seem to use forward models. When given the motor signals supplied to the muscles, a forward model will yield the results, which may be expressed in terms of the final position of the limbs, or of the sensory inputs associated with that position. It has long been known that a copy - the efference copy - of the motor outputs from the brain is sent from the motor cortex to other parts of the brain. This appears to be used as the input to one or more forward models, which then yield the sensory input that the movement will produce. This can later be combined with the incoming sensory feedback, which will be noisy, to produce a more accurate estimate of the actual current state of the system. Perhaps more importantly, since the forward model computation can start as soon as the efference copy is received - which is before the movement starts - the output can serve as a prediction of the expected movement and sensory input. This prediction can be used in many ways. For example, the predicted movement can be compared to the intended movement, and any difference (arising from noise, or from a deficient inverse model) can be compensated for much faster (though less accurately) than if feedback had to be relied upon. In addition, it can enable the brain to distinguish between movements caused by voluntary action (where the actual sensory feedback will match the prediction), and movements caused or affected by external factors or forces (where there will be a mismatch); and it can also be used to cancel out any sensory disturbances caused by voluntary movements (such as head movements) that affect the inputs from sense organs. In many cases, the processing using these forward models can be localised and identified in the brain by modern techniques of brain imaging.

Outside the context of control, there are many indications that imagined situations may lead to brain activity similar enough to the brain activity caused by the real situations to be regarded as primitive models of them. Most of these studies involve sensory imagery alone (e.g. Behrmann, 2000) but some combine sensory imagery with some imagined action on the image - typically rotation - and find evidence of apparently appropriate activity in motor areas of the brain (e.g. Richter *et al*., 2000). Hesslow (2002) has gathered together a great deal of relevant data in support of what he calls the 'simulation hypothesis'. He advances three 'core assumptions':

> (I) *Simulation of actions*. We can activate pre-motor areas in the frontal lobes in a way that resembles activity during a normal action but does not cause any overt movement.
>
> (2) *Simulation of perception*. Imagining that one perceives something is essentially the same as actually perceiving it, but the perceptual activity is generated by the brain itself rather than by external stimuli.
>
> (3) *Anticipation.* There are associative mechanisms that enable both behavioural and perceptual activity to elicit other perceptual activity in the

> sensory areas of the brain. Most importantly, a simulated action can elicit perceptual activity that resembles the activity that would have occurred if the action had actually been performed (Hesslow,2002).

He is able to present some neuroscientific evidence in support of the first two assumptions. If evidence could be found to support the third, it would supply a mechanism to support Dennett's speculations about the use of models in intelligent behaviour, because, as Hesslow notes, 'Once the mechanism of anticipation is in place, there is nothing to prevent the appearance of long chains of simulated responses and perceptions.'

The Possible Relationships Between Internal Models and Consciousness

There are many indications of the possible involvement of models in consciousness. The first is that conscious experience, especially when problem solving, sometimes seems to involve some sort of model of a real-world object - for example, when imagining a room in your house to see if a piece of furniture in a shop will fit. These 'mental models' have particular characteristics, and both these and the ways in which they can be used to solve problems have been studied in some depth by psychologists such as Johnson-Laird (1983). Johnson-Laird takes his inspiration from Craik, who wrote (at a time when intelligence and consciousness were regarded as inseparable):

> the nervous system is . . . a calculating machine capable of modelling or paralleling external events' (Craik, 1943, p. 120); If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to tryout various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the future, and in every way to react in a much fuller, safer and more competent manner to the emergencies which face it (Craik, 1943, p.61).

A second indication of the involvement of models comes from the spectacular examples of internal modelling recently made available by

Ramachandran and his collaborators (Ramachandran and Blakeslee, 1998). They involve so-called 'phantom limbs' - the internal image-like experienced representations of limbs that are absent, either through amputation or through developmental failure. In some cases, phantom limbs give rise to chronic pain or discomfort because the individual has lost the capacity to 'move' a phantom limb, and it has ended up in an awkward position. By using cunning experimental arrangements to replace the visual feedback that used to come from the now-absent limb with stimulation coming from a mirror-image of the remaining limb, Ramachandran has enabled some patients to regain their control of the phantom limb, and to move it to a comfortable position. This shows that the internal model is in some way affected by input from the real limb. On the other hand, a few rare cases reveal that phantom limbs can be present even in the congenital absence of limbs, showing that the model may be innate rather than learned. It is often difficult to make sense of these reports, which vary considerably, but the strong conscious awareness of one or more internal models of the disposition of the limbs seems beyond doubt.

A third line of thought suggests that the propensity of the brain for modelling may lead directly to consciousness itself, rather than merely providing the 'contents' of consciousness. In 1976 Dawkins set out his view that animals can be regarded as machines designed to ensure the survival and propagation of their genes, and remarked:

> Survival machines that can simulate the future are one jump ahead of survival machines who can only learn in the basis of overt trial and error. . . . The evolution of the capacity to simulate seems to have culminated in subjective consciousness. . . . Perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself (Dawkins, 1976).

This intriguing speculation seems to have been rather neglected; it is occasionally mentioned in the literature, but there does not seem to be any body of work directly inspired by it. Earlier, Minsky (1968) had identified some of the problems of thinking about consciousness as being caused by attempts at 'explaining the complicated interactions between the parts of the self-model', and had discussed systems containing both world-models and self-models, but had not proposed that the possession of a self-model might in itself give rise to consciousness. Perlis, in a series of papers summarised in

Perlis (1997), comes closer to Dawkins with his 'amazing-structures-and-processes paradigm', in which a self-modelling process enabling what he calls 'strong-self-reference' somehow gives rise to conscious experience. His work is particularly interesting in the present context because he explicitly considers the possible role of robots in shedding light on consciousness:

> I think that only by careful examination of human behaviour and the design of smarter robots will we be able to position ourselves to have more than merely prejudicial intuitions. . . . No-one, to my knowledge, has built, or even tried to build, strongly-self-referring machines. This in large part is simply due to the fact that no one has tried to build robots that can do very much reasoning, or even that can do very much common-sensical self-protection in a complex world. But strong selfreference is what an intelligent robot needs. . . (Perl is, 1997).

More recently, Damasio (1999) has proposed a neurologically based theory of consciousness in which the development of a primitive body-centred self structure plays a crucial role. The theory itself is complex, but his hypothesis is well summarised by Churchland (2002) in a paper examining self-representation in nervous systems:

> . . . the self/nonself distinction, though originally designed to support coherencing, is ultimately responsible for consciousness. According to this view, a brain whose wiring enables it to distinguish between inner-world representations and outerworld representations and to build a metarepresentational model of the relation between outer and inner entities is a brain enjoying some degree of consciousness. . . . Conceivably, as wiring modifications enable increasingly sophisticated simulation and deliberation, the self-representational apparatus becomes correspondingly more elaborate, and therewith the self/not-self apparatus. On this hypothesis, the degrees or levels of conscious awareness are upgraded in tandem with the self-representational upgrades (Churchland, 2002, p. 310).

The important point here is that it is not just the existence of some kind of self-representation that is held to lead to consciousness - it is the self-representation's relation to representations of things in the world, an idea that is more specific than Dawkins'.

An Incremental Programme

Our plan is to build a succession of robots that deal with the world by building and exploiting internal models; each new robot will be derived from, and capable of dealing with the world more intelligently than, its predecessor. Broadly speaking, there are two possible methods of doing this. The first would begin with a simple robot, embedded in a human-like world, and would attempt to increase its ability to cope with this world step by step; the second would move in stages from simple environments through progressively more challenging ones, adding the necessary extra abilities at each stage. Brooks (1991) has stressed the virtues of the incremental development of robots in the final target environment, and the pitfalls of beginning work in artificially simpler situations and hoping to migrate to more complex worlds. While we endorse his comments, we believe that the second strategy is more appropriate for this project; the reason is that we need to understand the situations and contingencies for which modelling may deliver benefits, and so we need to be able to control the exposure of the robot to those situations and contingencies. Our primary concern is not merely with building a robot that can cope, but also in understanding how modelling can enable it to cope; we therefore need to control what is present to be modelled, and what contingencies enable what kinds of modelling to deliver benefits. However, we recognise that the risks pointed out by Brooks may be inherent in this choice, and that a system able to cope with a simple situation may be wholly unable to cope when the situation is made even slightly more complex.

The path we propose to take is to begin with a very simple robot with a simple lifestyle (or mission) in a simple - but real- world. We will give it the ability to cope with that world, and will then progressively increase the complexity of its world and its lifestyle, and give it the bodily and computational resources to deal with the new problems. This is close to recapitulating evolution, but we do not commit ourselves to using some form of artificial evolution as the only mechanism, or even at all. Design is fast, comprehensible, modifiable, and extensible, and is what engineers are good at. Evolution is slow, intractable for most physical systems, and is not guaranteed to be comprehensible - indeed, there would be little point in evolving consciousness a second time, because the likelihood is that we would then be faced with a second type of system manifesting a consciousness that we could not guarantee to understand. In pursuit of the goal of producing a system that can be understood, we must at all times ensure that our systems are as transparent as possible - that we can see how they work, and in particular that we can see what their internal models are like, and how they deliver benefits.

Why use a real robot when using a simulated robot would be faster and more convenient? Parts of this question were answered very well by Brooks in a series of papers, including Brooks (1991). For example, using a real robot forces the use of the real world, and that prevents the avoidance, deliberate or unintentional, of the difficult problems inherent in the constraints of real world physics and real time. As far as convenience goes, it might well be easier to simulate a robot than to build it, but, as Brooks remarked, it would be much more difficult to simulate the rest of the world than simply to use it. Perhaps most importantly from the perspective of this project, our human consciousness developed as a result of evolution in the real world, and so if we were to succeed in developing some consciousness-like phenomena in a real robot, it would be reasonable to compare it with our own consciousness. If a modelled entity in a modelled world showed some consciousness-like phenomena, a critic could argue that any connection with our own consciousness would be formal or analogical rather than real: the consciousness of a simulated entity could only ever be a simulated consciousness. However, these considerations do not necessarily rule out the use of simulation in situations where the robot and environment are simple enough to give confidence that the simulation will be adequate, or when all that is required is an indication of whether a particular path is worth pursuing or not.

Many aspects of our programme are similar to those being followed by Edelman in his series of experiments with the Darwin robots. Edelman and his collaborators are attempting to understand the operation of the mammalian brain (Edelman, 1987) by using progressively more complete simulations of the brain to control a succession of progressively more complex robots in progressively more complex environments (e.g. Krichmar and Edelman, 2002). Our aim is to understand how the use of modelling to increase the capacity for intelligent behaviour may produce phenomena similar to those associated with consciousness. However great the difference in aims, in practice we will be using very similar equipment (robots, digital computers) and a very similar incremental strategy, and much of what he has said about the rationale and conduct of his decade-old programme will also apply to ours.

The Starting Point

The obvious way to start our programme was to place a simple robotic system in a simple environment, and to implement a simple modelling scheme to demonstrate the four uses of internal models. The aim was to create a system with the key fundamental attributes of any control scheme based on internal models; this could then serve as a baseline level of both structure and performance, from which we could pursue our strategy of incremental development. We did not expect to learn anything about consciousness from constructing such a system, and we were sure that its simplicity would deter others from over-interpreting any aspects of the system or its behaviour.

We began by looking for a suitable modelling scheme. An early candidate was the neural network implementation of a dynamical systems approach developed by Tani (1998). This was all the more attractive because it formed the basis of what we believe to be the first serious investigation of consciousness-related emergent phenomena using a robot as an investigative tool. For example, one of Tani's conclusions was that 'There is an essential structure of the "self' in the system and occurrences of "self-consciousness" are explained in terms of unfolding of this structure in time' (Tani, 1998). However, we felt that Tani's system was more complex than necessary for our purposes, and did not offer sufficient transparency, so we eventually settled on ARA VQ (Adaptive Resource Allocating Vector Quantization), a technique developed by Linaker and Niklasson (2000a,b) for dealing with abstract sensory flow representations.

For many reasons, it is useful for a robot to store sequences of past sensory inputs, or 'experiences', but memory limitations usually mean that relatively few experiences can be stored; compressing the data enables more instances to be stored, but not in a form that is immediately useful. Linaker and Niklasson attacked these problems by designing a method that allowed 'sensory flows' - raw temporal sequences of sensor inputs - to be encoded economically in a useful form. They had observed that when a simple robot is carrying out a simple task in a typically simple environment, the robot's sensory inputs (and also its motor outputs) are in practice often relatively stable for long periods of time. For example, when a robot is following a wall (a typical and much studied low-level robotic task), both its distance from the wall and its speed will be fairly constant, and so the sensor inputs from the robot's rangefinders, and of course its motor output commands, will also tend to be constant. If the robot then enters a narrow corridor, a typical control program will cause it to follow a course central to the corridor, often at a reduced speed; again, the sensor inputs (from both sides of the robot) and the motor output commands will tend to be constant, but different from those in the previous wall-following situation.

Using a simple robot simulation, Linaker and Niklasson began by regularly sampling each input configuration - the instantaneous sensory input and motor output - and developed a simple on-line clustering algorithm, ARA VQ, to detect and describe the small number of relatively stable and distinct input configurations, which they called 'concepts'. [1] Their robot could store long sequences of experiences very economically simply by labelling the different concepts, and recording the number of times each concept was repeated consecutively. For example, a run in which the robot followed a wall (concept W) for 29 time intervals, rounded a slow bend (concept B) for 4 time intervals, and then moved along a corridor (concept C) for 37 time intervals would reduce to (W,29; B,4; C,37).

The concepts have some interesting properties. Each one represents a range of input configurations which the algorithm has identified as forming a cluster. If a new configuration falls within that range, or close enough to it, it will be mapped to that concept, and the concept itself may be slightly modified to represent the new data rather better. However, if the new configuration falls far enough outside any existing concept, it may be used as the basis of a new concept; if more configurations close to the new concept are encountered, they will then be assimilated to it. We refer the reader to their papers (Linaker and Niklasson, 2000a,b) for a detailed description of their algorithm.

An attractive feature of this scheme is that it is possible to take a particular sequence of concepts recorded by the robot, and to use it to construct a step-by step representation of the robot's 'experience' and 'behaviour'. Each distinct concept can be characterised by the average reading of each of the sensor inputs and motor outputs, and so for each occurrence of

[1] We are unhappy with their terminology, but we retain it here for compatibility with their work.

a concept, it is possible to construct an approximate representation of (a) what must have been in the environment to activate the rangefinders, and (b) how the robot must have moved during the sample period. By plotting these movements step by step, and at the same time tracing out the points corresponding to the rangefinding data - a process Linaker and Niklasson call 'inversion' - it is possible to produce a representation of the robot's implicit model of its movement through the environment, and of its sensory contact with environmental features. The result looks like a map of the environment, and enables a quick appreciation of how well the implicit model corresponds to the real environment. Examples of the results of inversion are shown and discussed below.

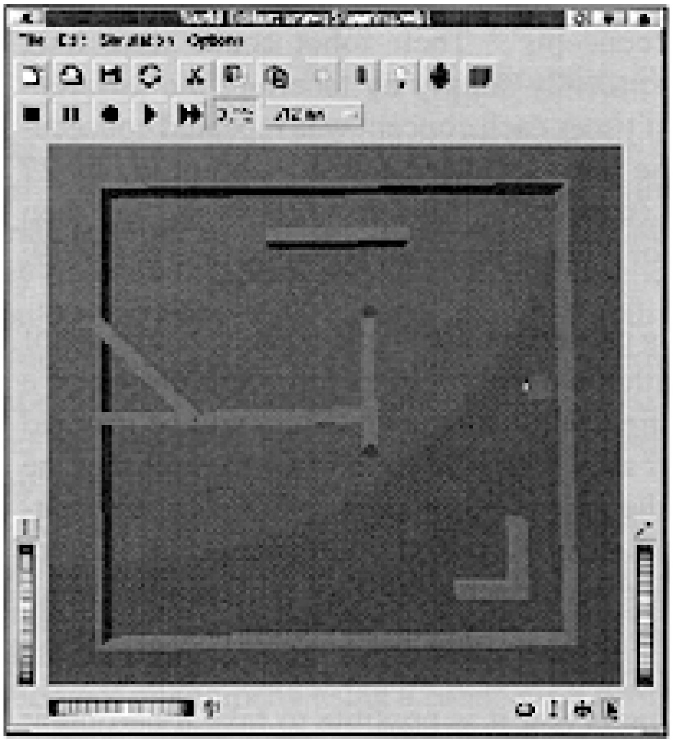

Figure 1. A screenshot of the environment and Khepera modelled in the Webots simulator. (Copyright CyberBotics Sari)

Linaker and Niklasson used a simulation of the popular Khepera robot (Cyberbotics, 2003b). The basic Khepera, although it is a complex and beautifully engineered product, is conceptually very simple: it is circular, with two independently driven wheels on each end of a diameter, and with six equally spaced rangefinders mounted on one semicircumference (facing forwards) and two on the other, facing backwards. Our first step was to reimplement their work using Webots, the very sophisticated and accurate simulation system provided by the Khepera's manufacturers (Cyberbotics, 2003a,b). Figure I shows a screen shot from the simulation of our first experimental environment. The simulated Khepera is programmed with a simple wall-following routine, and moves forward keeping a set distance from the wall on its right while avoiding collisions. We found that ARAVQ rapidly built a stable set of concepts, and that it was easy to identify these concepts with particular sections and features of the environment.

In Figure 2 we show how the process of inversion operates. The image on the left represents the inversion of two successive activations of the same concept from a Khepera moving along a corridor. (The size of the step between the activations has been scaled up for clarity.) The Khepera is represented by a circle, with the front-back and left-right diameters drawn in, and with a triangle superimposed on the left-right diameter to show the orientation. The positions of the rangefinders are indicated by the small circles on the periphery of the robot. In the initial (lower) position, small triangles mark the points corresponding to the readings of the two leftmost and the two rightmost rangefinders, as encoded in the concept. In the upper position, the representation of the Khepera has been moved by an amount corresponding to the operation of the left and right motor outputs for one time step; since the outputs encoded in the concept are equal, the Khepera has simply been moved forwards. The new points defined by the rangefinder readings are marked, and lines are drawn joining the successive readings of the points defined by each rangefinder. These show the line of the corridor quite accurately.

The image on the right shows two successive activations of a concept formed by a Khepera turning in response to a corner. In the initial position, all the front-mounted rangefinders except the leftmost detect the right-hand wall of the corner; its approximate position can be seen by joining the rightmost triangles on each of the five curves traced out by the sensors.

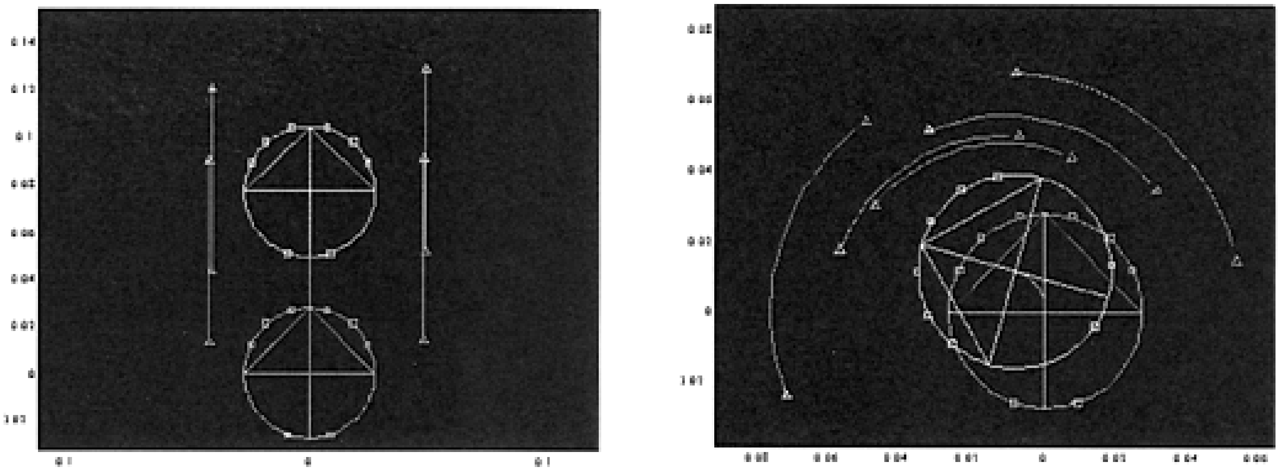

Figure 2. How a map of the environment can be reconstructed from the record of the robot's sequence.

In the record on the left, the 'corridor' concept is active for two successive time intervals. The model of the Khepera is displaced by an amount corresponding to the concept's motor activation. The sensor readings for the two leftmost and two rightmost sensors give the distance of the reflecting surfaces (walls) at each position; by joining successive corresponding points, we can see part of an environmental map. The record on the right shows a transition involving a small translation and a large rotation, with the derived map. (After Linaker and Niklasson, 2000a)

The Khepera's motion, as encoded in the concept and as illustrated by the change in position and orientation of the triangle on the diameter, is a left turn of almost 90° combined with a small forward movement. Since the concept is the same as at the previous step, the relative positions of the rangefinder distances are the same in relation to the Khepera; as can be seen by joining the leftmost triangles on each curve, they mark the approximate position of the other wall of the corner. Although the line of the corner is clear enough once one is used to the representation, it is less intuitive than the image corresponding to a wall or corridor.

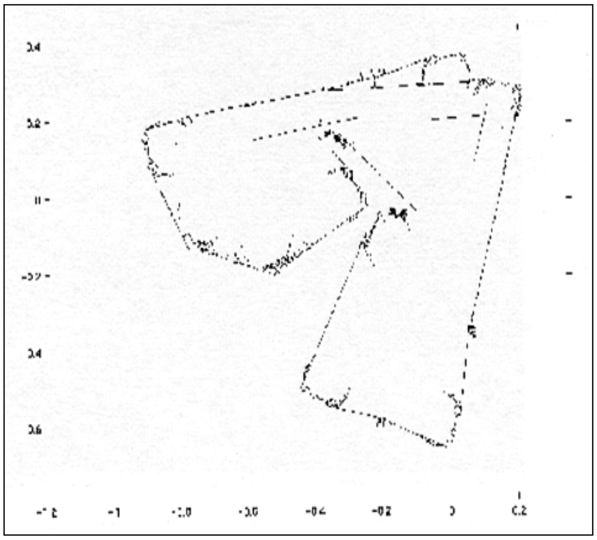

Figure 3. The environmental map derived from a simulated Khepera in the environment of Figure 1.

Although rotational inaccuracies distort the map, it contains good local representations.

In Figure 3 we can see the results of inverting the series of concepts activated by the simulated Khepera in making a circuit of the environment of Figure 1. The local representation of detail is surprisingly good, but the global map is distorted, primarily because of errors associated with rotational movement - a well known problem in robotics. (Robot simulations are often criticised because they are held to give insufficient weight to the realworld noise and random errors which affect the sensors and motors of real robots. In Webots, noise can be added to the simulated inputs and outputs to compensate for this.) In order to bring this work into the real world, our next step was to see if ARAVQ was capable of forming stable and potentially useful concepts in a real Khepera operating in a real environment. Perhaps because of the Khepera's lack of dynamics and simple sensory apparatus, we encountered no problems in making this transition. Figure 4 shows a typical environment being explored by a real Khepera programmed with the usual simple wall-following algorithm; Figure 5 shows the results of inverting the sequence of concept activations for a single circuit once the concepts had stabilised. As can be seen, the representation of the implicit model appears qualitatively similar to the results obtained in simulation, with the good local detail, but with the global representation distorted by rotational inaccuracies.

We have so far shown that a robot using the modelling scheme, ARA VQ, can build an internal model of its interaction with an environment, and that this internal model (comprising both the concepts and the list of the sequence of concepts encountered in the environment) can be interpreted to produce a comprehensible representation of the outcome of the interaction. We now need to show that the model can support the four functions of modelling identified above; however, it will be convenient to examine them in a different order.

Enabling and improving control

The Khepera had learned the concepts passively, while moving through the environment under the control of the wall-following program. In order to make the robot move under the active control of the concepts, it was clear that the basic requirement would be to select a concept corresponding to the existing stimulus configuration, and then execute the motor outputs represented by that concept. However, the motor outputs were themselves part of the concept - how could they be taken into account before they had been generated? Nathan Grey, who carried out the robot experiments, suggested an elegantly simple method: select the most appropriate concept solely on the basis of the rangefinder data, and then execute the motor outputs specified by that concept. The results at first surprised us. Under the control of the concepts, the robot produced smooth and accurate wall-following - a great improvement on that produced by the original wall-following program! On reflection, however, this was to be expected. The wall-following program had been made deliberately crude and jerky, responding to inputs with large changes in motor output, because we wanted to present ARA VQ with a reasonably difficult modelling problem. Since the motor output specified by a concept is an average of the motor commands that have been produced within that concept, it is bound to vary less than the outputs that contributed to the formation of the concept, and so local smoothing is inevitable. We had thus used one facet of the internal model-the concept - to enable control; by accident, it had also improved control. It should be noted that we disabled the concept formation during this test phase.

Processing novel or incomplete data through an existing model to produce 'appropriate' actions

When the robot is run in the mode described above, the incoming set of rangefinder readings will simply be mapped to the concept with the most similar set, as defined by a simple measure. Since there are six rangefinders, each taking readings independently, it is unlikely that exactly the same set of readings will be encountered twice in a single run; because of minute environmental variations, this would be true even if the rangefinders behaved perfectly, but the certainty of sensor error (especially occasional failed readings) makes it something to be expected anyway. Because of the simplicity and regularity of the environment, and the relatively weak demands of the control task (wall-following), the selection of the concept with the most similar input set did in fact produce appropriate actions from novel or incomplete data.

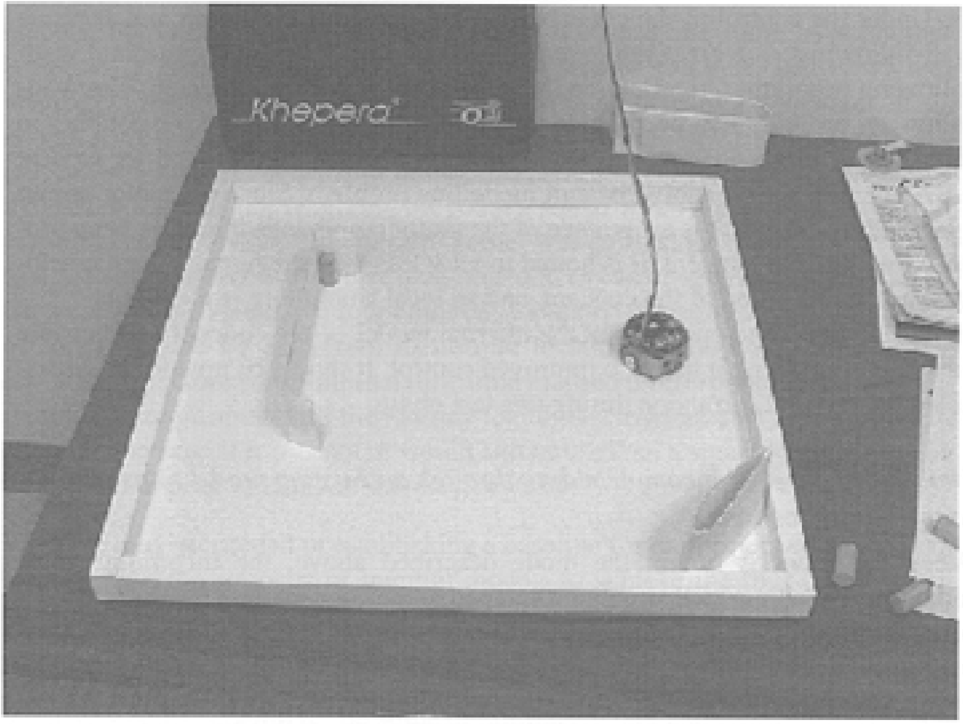

Figure 4. The Khepera robot is modelling the environment using ARAVQ

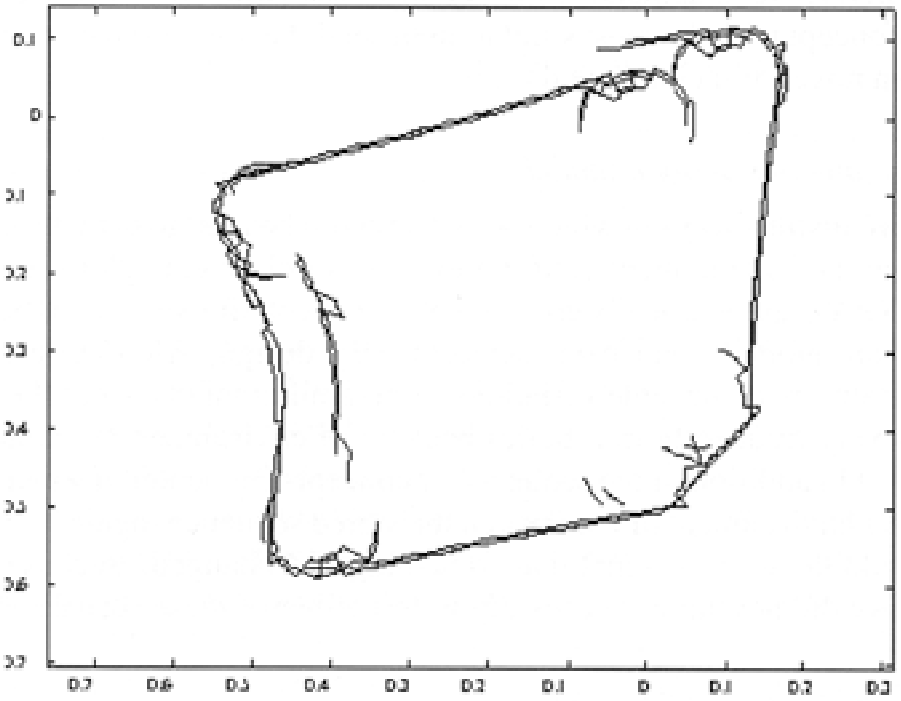

Fig 5. The environmental map derived from the modeling process of Figure 4.

Although the global map is distorted, the local detail is excellent.

Detecting anomalies using a model

There are two distinct ways in which ARA VQ could be used to detect anomalies. If a persistent and stable input configuration is not close enough to an existing concept, ARA VQ automatically creates a new concept; this step could be used to signal that an anomaly had been detected. (By design, ARA VQ effectively ignores momentary or unstable variations from familiar inputs, since these situations occur very frequently in robotics because of environmental variation and sensor noise.) In addition, if the sequence of concepts encountered when traversing a familiar environment did not match the stored sequence (again, within limits), this could be used to signal that something had changed. Because of their simplicity, we did not think it necessary to test either of these scenarios.

Using a model to inform decisions

Since the wall-following task (following a wall on the right) produced fairly invariant linear paths around the environment, there was not a lot of scope for presenting the robot with a decision to be made on the basis of its internal models, because it could do nothing about choosing one route over another. However, it could at least decide whether to stop or continue, and so we took inspiration from the behaviour of a prey animal on detecting a predator. If the animal is close enough to home to make a successful run for it, it will do so; if it is too far away, it may freeze into immobility, and so avoid detection. We played the part of the predator by giving the robot a signal while it patrolled the periphery of the environment. The robot's 'home' was defined as a short corridor section with a distinct concept corresponding to it, as can be seen in Figures 4 and 5. On receipt of the signal, the robot internally stepped through the stored sequence of concepts, one by one, from its current location. If the home concept was encountered before some internal time limit elapsed (corresponding to internally stepping through about half the journey round the environment) the robot would continue moving; if it was not, it would halt until given a signal to proceed. This was trivial to program, and of course it worked.

We have thus succeeded in establishing a baseline for our project - a robot that uses perhaps the simplest of internal models to achieve the four functions we have identified as being of interest. We have also made the internal models transparent to some degree. There is, we hope, no sign of anything to which anyone might want to apply the adjective 'conscious', although the system may perhaps qualify as being 'cognitive' on some definitions (Clark and Grush, 1999). There is nothing that we would wish to characterise as any sort of self-model, although the details of the robot's embodiment are clearly crucial in determining the form of the representations of the internal models. The only way is up.

Increasing the Complexity of the Robot

The nature of the robot's sensors obviously places some limitations on the nature and extent of the information the robot can acquire about the environment and its contents. The same is true of animals, as von Uexkull recognised many years ago (von Uexkull, 1934). A robot fitted only with horizontal range sensors would never be able to discriminate between a low structure and a high one with the same size and shape at the base, however important the distinction might be for its well being or survival. On the other hand, an extra set of range sensors mounted high on the robot would be utterly useless if the environment contained nothing that was able to stimulate them. Somewhere in between is the case where the robot can acquire information about some aspect of the environment, but there is no contingency where that information can make any difference to the robot's performance. In a given environment, therefore, the sensory provision should enable information to be extracted and used (perhaps in a modelling process) to provide benefits at least equalling the cost of that extraction and use. It will not be necessary to provide an improved capacity for sensing every time the environment is changed; the vertebrate eye has remained essentially the same for a very long time - it is the capacity of the brain for processing the information from the eye that has changed.

As regards effectors, it is worth noting that the majority of research robots can do nothing except move through the environment. This is not solely due to the lack of any grasping or manipulating appendages, because many robots are potentially able to act on their environment and change it without any specialised effectors, simply by pushing things; the problem is simply that the experimental environments rarely contain anything that can be pushed! Actions more complex than pushing require some specialised effectors, and these in turn require the presence in the environment of suitable structures or objects upon which they can act. In the context of our robots, or of animals, there is a further constraint very similar to that noted for the sensory equipment: acting on the environment should be able to bring some benefit, otherwise the provision of the effectors represents nothing more than a cost. Of course, effectors, especially if they have redundant degrees of freedom, may require modelling for control purposes. In addition, effectors enabling the manipulation of an object can yield information about the object that would not be obtainable in any other way - for example, how to manipulate the object, as well as how the object can be manipulated. Finally, any addition to an agent's complement of effectors is likely to increase its potential for engaging the environment in different ways, and this will enlarge the set of affordances offered by the various environmental features, again adding to the complexity of the system as a whole.

The state of a typical mobile robot such as the basic version of the Khepera is described by its pose alone - its position and orientation - simply because the body is rigid. However, once there are movable body parts or effectors, especially if they are articulated, their dispositions may need to be taken into account by the control system. If these body parts or effectors carry or serve any sensory apparatus, then their movements and positions may also be useful to the sensory system; some examples of the use of such information were given in the section above on the use of models in the brain.

The three factors mentioned so far - sensors, effectors, and the body would serve to define the non-behavioural complexity of most present-day robots, the general idea being that the complexity of a robot will increase with increasing complexity in its sensors, effectors, and body. However, these factors would be inadequate to describe most animals, because there is a further determinant of complexity to take into account. Since animals are biological entities, they are subject to the biological constraints of metabolism, growth, and ageing, and these complicate matters. The most appropriate current action for an animal depends on the state of variables such as hunger, thirst, fatigue, on whether it is a juvenile or an adult, and so on; the only corresponding factor in most robots is the state of their battery, though robots with artificial metabolisms are beginning to appear (Kelly et at., 1999; Wilkinson, 2000). Many of the shifting multiple goals that animals have to deal with come from these biological characteristics, and we think it likely that a robot with the animal-like intelligence that might lead to consciousness would have to be capable of dealing with similar time-varying, cyclic, and historical constraints.

Increasing the Complexity of the Environment

There are many aspects to a robot's or animal's environment. It is both a space, and a range of objects and materials contained and localised within that space. The techniques used by animals to move through space, and the extent to which they use modelling, are becoming increasingly well understood. Paradoxically, some of the most difficult environments spatially are those that appear least complex, such as the Sahara desert; environments rich in spatial information are relatively easy for both animals and robots to deal with. What makes an environment challenging and complex is the nature and disposition of its contents. If we take the baseline environment as being the simple fixed and bounded space of the baseline experiment, it is possible to see how the introduction of various types of objects will require the robot to improve its abilities, especially where modelling is concerned.

The first step might be the introduction of moving objects. These will force the robot to avoid them in real time; distal sensors will be necessary for this, and in a cluttered environment the ability to predict an object's trajectory may prove necessary. The Khepera's sensors are inadequate for this task, and so it is likely that the next set of experiments will require a new robot fitted with long-range ultrasonic or laser rangefinders, or a suitable vision system. The next step could be the introduction of a variety of objects with different values for the robot, perhaps corresponding to food values and levels of toxicity; the relationship between the sensory characteristics of the objects and their value will have to be learned. (This is the stage at which Edelman is currently working - see Krichmar and Edelman, 2002). If objects of different types tend to occur in particular parts of the environment under particular circumstances, advantage will be gained by modelling and exploiting this. If the value of an object can only be obtained by a certain sequence of actions (as in cracking a nut), rather than by simply gripping it as in Edelman's scheme, the problem posed by the environment becomes even more difficult.

Passive objects such as sticks, stones, fruit, and plants are clearly important to many animals, and to humans, but the really challenging components of most animals' environments are active objects, or agents. These may be of other species, and may be parasites, prey, predators, and so on. In dealing with any of these agents, reactive strategies will inevitably be outperformed by predictive strategies, and these will require some form of modelling. However, the most interesting agents to introduce into the robot's environment will be robotic analogues of con specifics, which may be competitors, collaborators, or mates, or even all three at different times. Where there are repeated contacts with the same individuals, and where there are individual differences, it may become necessary to be able to identify individuals, and to remember their characteristics and the outcomes of previous interactions. Of course, we will now be at the level at which various theories have proposed the origination of social intelligence, communication, language, and consciousness; the important thing is that we will have reached it in an incremental way, and will have a complete know ledge and agood understanding of the modelling and other resources available to the robot enabling it to cope with the environment before each new introduction.

Effective Choice and Planning: Could, Should and Would

Although we distinguished four possible roles for modelling in a previous section, it is clear that one, the use of modelling to inform decisions, is more complex than the rest, and capable of much greater development. This is not the place for a full analysis of the system features and capabilities required to implement and support such a scheme; indeed, to the best of our knowledge, there is as yet no such analysis in the literature. However, some aspects of what must be capable of being done are particularly relevant to the development of consciousness, and we describe them briefly below. They can easily be organised under three headings: what the system could do, what it should do, and what it would do.

*What the system **could** do*

For effective planning, the internal models used by the system must reflect the real world with sufficient accuracy. There are two important aspects to this. First, the modelled properties of the physical world - objects and space - must reflect the real properties sufficiently accurately. Second, the abilities of the system to act on and in the physical world must also be modelled well enough, and this means modelling the physical properties of the body, the ability to control it, and also the ability to process information appropriately. As Ramachandran and Blakeslee note:

> It is always obvious to you that there are some things you can do and others you cannot given the constraints of your body and of the external world. (You know you can't lift a truck. . .) Somewhere in your brain there are representations of all these possibilities, and the systems that plan commands. . . need to be aware of this distinction between things they can and cannot command you to do. . . . To achieve all this, I need to have in my brain not only a representation of the world and various objects in it but also a representation of myself, including my own body within that representation. . . . In addition, the representation of the external object has to interact with my self-representation. . . (Ramachandran and Blakeslee, 1998, p. 249).

Ramachandran's and Blakeslee's comment about the interaction between the self- and the object-representation is particularly interesting, and worth exploring more deeply. If a system is modelling two objects interacting - for example, two billiard balls colliding - then one obvious way to organise this is to allow separate instances of models of each object to interact, rather than to have a dedicated integrated model of the two objects interacting under all possible circumstances. (Incidentally, there is plenty of evidence that humans and animals are supplied with innate knowledge of a wide range of object properties, including their behaviour in collisions - e.g. see Hauser, 2000.) Although the most obvious models available to the system would be those based on perceptual information, as suggested by Hesslow (2002), any other adequate representation would do. When the situation being modelled is the body interacting with an object, the same approach of allowing models to interact seems appropriate, especially as Ramachandran and Blakeslee present evidence for an innate model of at least some characteristics of the physical body, namely the limbs. However, there is a difference between a model of the body and a model of an object such as a billiard ball: the interactions of a body model with an object model as do not just depend on the modelled physical properties of the two entities, but depend also on how the body model is controlled, and this in turn depends on the information available for controlling the body model. Modelling the catching of a ball requires the modelling of the trajectory of the ball, the modelling of the perception of the trajectory, and the modelling of the arm and hand movements produced by control systems using data from that perception. This suggests that object models based on perceptual data might be particularly useful for body-object interactions.

In this context, it may be worth pointing out that the development of virtual reality technologies has provided a powerful metaphor for many aspects of consciousness, one that is used by Metzinger (2000) and Dawkins (1998), among others. Virtual reality provides an individual with sensory inputs - chiefly vision, touch, and sound - which match the inputs that would be produced by the physical environment being modelled if the individual were at a particular location in that environment. The physics of the environment - the positions and masses of objects, the way balls bounce, the way light behaves - may be modelled and used in the underlying computations, but what is presented to the individual is not a three dimensional physical model of the environment in x, y, and z coordinates, but a simulacrum of the view from a point in such a model. The visual imagination also seems to use such a centred view, and this may indicate that at least some of the models in the brain are similar to the products of virtual reality. However, Grush (2002) has clearly set out the differences between purely sensor-based modelling (modal emulation, in his terminology) and modelling which deals with what he calls the egocentric space/object environment (amodal emulation); as he makes clear, visual imagery may well involve both types of systems.

*What the system **should** do*

If evolution has done its job perfectly, then out of all the actions or sequences of actions that an animal could perform, the one it should perform will be the one that can be expected to make the highest marginal contribution to the propagation of copies of its genes. There is clearly no credible way for any animal to calculate this exactly at every instant; some approximate computation, or implicit computation, must be carried out instead. In some cases, however, the result of such computations can be uncannily accurate. For example, when life's problems are simplified by a specific situation - feeding young - so that maximising the rate of acquisition of energy from food is all that matters, some birds can make decisions about when to gather food at particular time-dependent food sources (tidal mud, rubbish dumps) that appear to be optimal.

Methods for the generation and evaluation of plans, by algorithmic and heuristic means, have been at the centre of conventional AI since the discipline was founded. Unfortunately, there is little or no useful information about either the representation of sequences of actions in the brain, or the mechanisms by which sequences of high utility are generated rather than those of low utility. However, there is a wealth of evidence that the encoding of utility in the brain is strongly associated with emotion. Rolls (1999) examines the issue at length, and suggests that the notions of reward and punishment provide a unifying structure - a common currency - with plausible links to the mechanisms of choice and multi-step planning. 'Is there any alternative to such a reward/punishment based system in this evolution-by-natural-selection situation? I am not clear that there is' (Rolls, 1999, p. 273). Edelman is also convinced of the necessity for such a value system, and again chooses a scheme with a positive and a negative aspect based on appetitive and aversive events (Sporns et al., 2000). The implication of such schemes is that the action sequences generated by an animal are those that either maximise the expected reward, minimise the expected punishment, or produce the most positive net effect. The usefulness of such decision-making is determined by the correspondence between those decisions and those which would be made by some ideal system using values derived from the actual expected reproductive benefits. We propose to adopt such a system in our project. Of course, we do not mean to imply that there is nothing more to emotion than the production of a reward or punishment signal. Strong emotions may trigger behaviour such as fighting or flight, or cause bodily changes preparing the organism for such actions. In social species, the expression of emotion plays a key role in social interaction. Human consciousness is saturated with feelings and emotions varying on many more dimensions than mere sign and strength; we do not know why this is so, but its reality is unquestionable. However, from a functional point of view, we need some method for the comparative evaluation of action sequences in order to achieve the selection of the most advantageous, and the assessment of each sequence and each component action along a single positive-negative dimension is both simple and apparently biologically plausible.

The assumption made explicitly by ourselves, and sometimes implicitly by others, is that the generation of action sequences of high utility involves the sequential modelling of possible actions and of their objective effects, together with some evaluation of their benefit to the organism. Somehow, actions or action sequences evaluated highly tend to be executed, and good enough sequences are executed often enough to give some advantage over systems without this capacity. There may be some method of achieving all this without modelling, but we cannot at present imagine it. However, there is no point at all in modelling unless there is an evaluative component; unless the process leads to the agent doing what it *should* do, modelling will confer no benefit.

The idea that sequential modelling is at the root of planning was central to the early efforts in artificial intelligence. It soon became apparent that, in the paradigms used at the time, the key problem was not just to calculate the outcome of a particular sequence of actions, but to search among a number of possible sequences to find one that was acceptable. AI soon became centred around the exploration and evaluation of different search methods. It may well turn out that issues related to search will be important for the ideas presented here, but our immediate concern is with the nature of the sequential modelling itself.

*What the system **would** do*

Even if a system can identify the actions it is possible for it to carry out in various situations, and is able to run internal models of multi-step plans, and to estimate the rewards, punishments, and costs of each stage in the plan, that may not be the end of the story. Rolls believes that the human brain contains two independent methods of selecting action: the implicit system, also present in non-human primates, which selects the next action on the basis of 'assessment of the reinforcement-related value of a stimulus' (Rolls, 1999, p. 256); and the explicit system, a language-based long term planning system. He remarks of the hypothesised explicit planning system: *'The process may enable an available reward to be deferred for another reward that a particular multistep strategy could lead to'* (Rolls, 1999, p. 270). The implicit system is essentially for selecting the immediate action with the highest expected reward; the planning system will only confer benefit when it selects an action with a lower immediate reward (as judged by the implicit system) that will lead to an eventual higher net benefit. If Rolls' conjecture about 'dual routes to action' (Rolls, 1999, pp. 255-61) is correct, then for the planning system to operate successfully, it must in some way resolve any conflict with the implicit system in its own favour. It seems unlikely that this will be achieved by the planning system always completely suppressing the older and faster implicit system; we might expect to see a degree of flexibility, with the implicit system being more likely to win if the immediate reward is high and if the eventual reward favoured by the explicit system is remote or only slightly superior.

If this is the case, then it adds a new dimension to the formulation of a plan. To be effective, the planning system must be able to predict whether it will prevail over the implicit system at every stage before the final reward is reached. To do this, it will need an accurate model of the strength of the implicit system's activation in any given modelled situation, as well as an accurate prediction of the outcome of the conflict between the implicit and explicit system. This amounts to assessing what the system as a whole would do in a given situation. An example might be a recovering alcoholic's plan for buying cigarettes during a downpour. Rather than walking two hundred yards in the pouring rain to a tobacconist's, he might decide to get some from the bar next door. As a plan, it is good as far as it goes - except that the proximity of immediate sources of alcoholic reward in the bar may cause his implicit system to suppress the planned action of leaving immediately in favour of that of buying a drink, and he should consider this before committing to the plan.

An Emergent Architecture: The Agent-Model and the World-Model

Looking at the above material, it is clear that the usefulness of modelling does not depend only on the formation and exploitation of models of the external world, but also on the modelling of many aspects of the agent itself. The agent must model the characteristics of its body and its sensors, and its ability to control the body as a function of sensory inputs, and to operate on objects and move through the environment. It may have to model the reward and punishment expected from modelled actions and situations; it may also have to model the behaviour of its own decision systems. As engineers, it is our working hypothesis that the best way of organising such a system will be for all the models of the various aspects of the agent to be linked functionally into a single composite model, the internal agent-model, which will have a degree of separation from all the models of the external world forming the world-model (Figure 6). Planning will then involve an ongoing interaction between the agent-model and the world-model, with each affecting the other; the actions of the modelled body will affect the modelled objects in the world-model, and the changes in the modelled objects as a result of those actions will cause corresponding changes in the modelled perceptions of the agent-model. Information about planned sequences of action with high evaluations must somehow be passed to the part of the system responsible for initiating actions, and these sequences must be selected sufficiently often and in appropriate circumstances to produce net benefits from the planning system, and to outperform competing designs.

The agent-model can directly affect only itself and the world-model; it can be directly affected by the world model, and by the agent qualities it models - the body, the sensors, the evaluations of the consequences of actions, and so on. In order to function well, it must be promptly updated with current information about the agent, so that it can accurately model and predict the agent's characteristics and behaviour. If it is a good model, and if the system for translating preferred plans into action is accurate, then the agent-model's preferred plans will be executed - presumably after a delay. The predicted changes to the body will occur, and the agent-model will be updated as predicted; the predicted changes in the world will occur, the world-model will be updated, and the updated world-model will affect the agent-model. The agent-model's plan will have been executed, and all the information coming in to the agent-model will be as predicted. In other words, the agent-model's actions in its original trial interaction with the world-model will produce the predicted effects in the updated agent-model and world-model. This virtual agent, trapped in its virtual world, will therefore appear to be able to act in the real world without in fact being able to do so.

If this separation between world-model and agent-model turns out to be advantageous or necessary in robots, then it invites the speculation that a similar separation may have happened in animals and humans. In other words, the efficient exploitation by evolution of the modelling abilities of the vertebrate brain may have led to an agent-model and a world-model as described above. If this has happened, then we suspect that the human agent-model will turn out to be the structure supporting conscious experience, and so the characteristics of conscious experience will be determined by the characteristics of the agent-model, and experience of the world may correspond to the agent-model's interactions with the world-model. We call this the IAM (Internal Agent-Model) theory of consciousness. Although the agent-model could equally well have been called the self-model, we wished to avoid the associations created by the many uses of the word 'self' in the context of consciousness. For example, Strawson (1999) distinguishes seven distinct aspects of the sense of self 'in so far as the self is experienced specifically as an inner mental presence' (Strawson, 1999, pp. 490-1); Damasio (1999) identifies the proto-self, the neural self, the core self, and the autobiographical self; and so on. By using the concept of the agent-model, we avoid identifying it with any particular aspect of the conscious self. Interestingly enough, the philosopher Thomas Metzinger has proposed a theory of consciousness explicitly based around the concept of the self-model. Although his is primarily a phenomenological analysis, it is sometimes intriguingly close to the position we have taken:

> The phenomenal self is a virtual agent perceiving virtual objects in a virtual world . . . I think that 'virtual reality' is the best technological metaphor which is currently available as a source for generating new theoretical intuitions. . . heuristically the most interesting concept may be that of 'full immersion' (Metzinger, 2000).

He also notes that the phenomenal self-model

> is a plastic multimodal structure that is plausibly based on an innate and 'hardwired' model of the spatial properties of the system (e.g. a 'long-term body image'. . .) while being functionally rooted in elementary bioregulatory processes. . . .

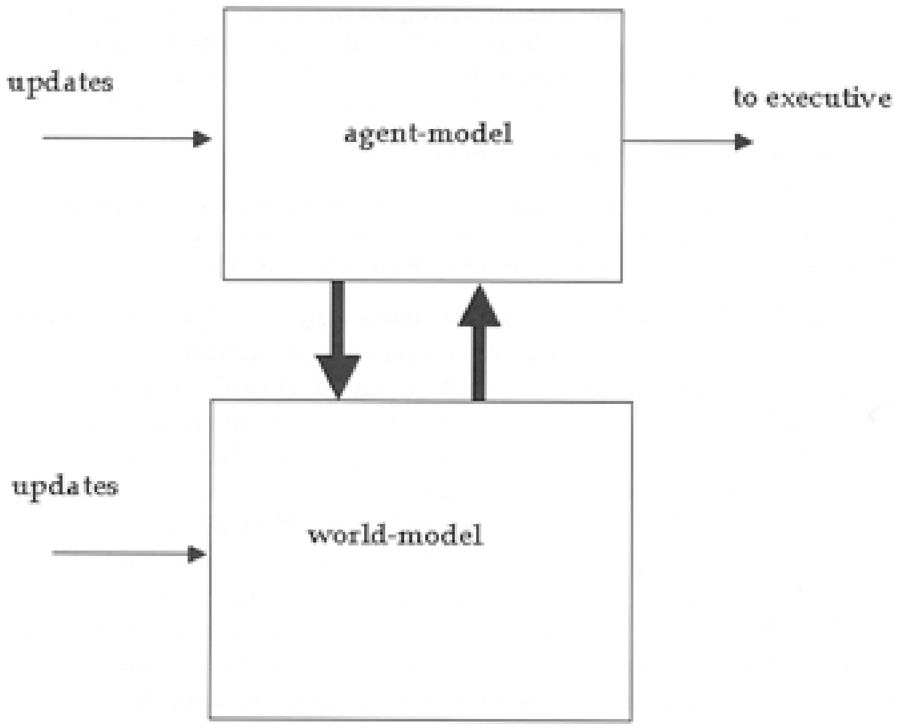

Figure 6. The hypothesised separation of the models involved in planning into two functional components, the agent-model and the world-model.

> The agent-model operates on the world-model, and is itself affected by changes in the world-model. It may also be updated by other inputs, particularly those involved in the evaluation of the consequences of action. The world-model can be manipulated by the agent-model, and is also updated by sensory inputs. Information about favoured sequences of actions is passed from the agent-model to the executive, where it may influence the actions actually selected.

The biologist Holk Cruse has also made an interesting contribution in this area (Cruse, 1999) and has ended up at a position very close to Metzinger's and to our own. His concern is with the possible structural and functional underpinnings of what he calls HIP systems (HIP = Having Internal Perspective), which he contrasts with NIP systems (Not having Internal Perspective). He begins by noting that the internal connections of simple feedforward neural network control systems 'can be interpreted as comprising an implicit world model (of that part of the world that is of importance for the system), because in some more or less indirect way it represents the properties of the world and the appropriate reactions' (Cruse, 1999). He notes that the inclusion of some recurrent (feedback) connections in such networks enables them effectively to predict sensory inputs, and to prepare the appropriate actions. He refers to both types of network (feedforward and feedback) as containing *non-manipulable* world models, in contrast to a third type of feedback network which can be isolated from its sensory inputs and motor outputs, and which can therefore 'play around in its virtual internal world' using what he calls its *manipulable* world model. Arguing that a world model that includes only 'information concerning the outer world' provides little support for a first-person internal perspective, he then proposes that things might be rather different in the case where 'some of the physical properties of the system itself are embedded into the internal world model' (Cruse, 1999). Drawing on some of his own previous work in neural networks for the control of movement (Steinktihler et al., 1995) he sketches out a design for a system which contains a manipulable model of its own body. Finally, he advances his hypothesis:

> a system comprises a HIP system, i.e., has a first-person perspective, when (a) it contains a manipulable world model that includes properties of the system's own body which (b) can be used to compare the ('virtual') data of this model with those provided by the 'real' data from the sensory input. . . (Cruse, 1999).

Although his primary concern is clearly with the problem of natural consciousness, he does note that, if the hypothesis is correct, 'artificial systems could be constructed such as to have first-person perspective, i.e., to become HIP systems' (Cruse, 1999).

A number of recent papers and books have proposed the involvement in consciousness of some low-level neural structures representing and bringing together the physical, metabolic, and emotional state of the organism. In his review of Damasio's *The Feeling of What Happens* (Damasio, 1999), Douglas Watt summarised Damasio's position:

> consciousness requires that the brain must represent not just the object, not just a basic self structure, but the interaction of the two.

He went on to note:

> This is still an atypical foundation for a theory of consciousness, given that until recently, it was implicitly assumed that the self could be left out of the equation. There has been a recent sea change on this crucial point. . . (Watt, 2000).

A collection of papers on the self taken from several issues of the *Journal of Consciousness Studies* (Gallagher and Shear, 1999) gives a wide-ranging set of views on the nature of the self. Most are centred around the characteristics of the self as revealed through phenomenological analysis, but some regard the self as a structure or process, underpinned by neural systems, that provides some functional benefit to the organism as well as underpinning consciousness. Panksepp (1998) is an example:

> I advocate the position that the roots of the self go back to specific. . . sensory-motor action circuits within the mammalian brain which can generate a primitive sort of intentionality. . . and primitive forms of psychic coherence. . . by interacting with various emotional and attentional circuits that encode basic biological values. . . . These interacting circuits have specific neurochemical codes that may generate distinct types of neurodynamics within primitive core systems of self-representation that first symbolized organisms as coherently active creatures in the world (Panksepp, 1998).

Our proposal differs from the schemes of Panksepp, Metzinger, and Damasio in that the agent-model does not grow out of some primitive protoself or neural structure, but is instead a technical requirement of the very high level task of planning. Although their formulations are seductive, and elements of them could be accommodated within our scheme, we believe it will be best to pursue our approach in isolation so that it will be clear which properties of consciousness, if any, may derive exclusively from the planning requirement.

The Signs of Consciousness

Even if we succeeded in producing a robot system which successfully used an agent-model and a world-model for planning, how could we set about analysing the system in terms of possible links with consciousness? If we could observe the changes in the 'contents' of processes within the agent-model, and the relationship between those contents, the real world, and the actions of the robot, then we would be able to attempt to compare those characteristics of the agent-model with the characteristics of human consciousness: Baars' Global Workspace Theory might be a suitable framework to use (Baars, 1988). If they turned out in the best possible case - to be practically identical, what would we be able to say? It would not be possible to claim that we had produced machine consciousness in the sense of Block's P-consciousness - we see no prospect of that claim ever being verifiable. It might be reasonable to claim, however, that we had identified the functional origins and components of the architecture within which consciousness exists, and that we had built a system with such an architecture. We would certainly be happy with this. When consciousness is operating normally, there appears to be a close correspondence between subjective experience and reality. However, as is clear from examining the session headings in any modern consciousness conference, we now know that much of this correspondence is illusory. As Norretranders puts it:

> Consciousness is a peculiar phenomenon. It is riddled with deceit and self-deception; there can be consciousness of something we were sure had been erased by an anaesthetic; the conscious I is happy to lie up hill and down dale to achieve a rational explanation for what the body is up to; sensory perception is the result of a devious relocation of sensory input in time; when the consciousness thinks it determines to act, the brain is already working on it; there appears to be more than one version of consciousness present in the brain; our conscious awareness contains almost no information but is perceived as if it were vastly rich in information. Consciousness is peculiar (Norretranders, 1998, p. 286).

If the characteristics of the information flows through the agent-model in our system turned out to be peculiar in the same ways without our having deliberately made them so, then it would strengthen the claim that consciousness was the outcome of the operation of a similar system, and that the robot system might therefore represent a form of machine consciousness. Unfortunately, we do not believe that it is realistic to expect such high levels of performance from a robotic system in the near future, and so we must begin by looking for signs of more basic and non-illusory attributes of consciousness as set out by Baars.

Of course, engineers, like doctors, are familiar with the idea that a failing system can often reveal more about its principles of operation than can a system working normally. If you asked ten competent engineers to build a device conforming to some specification, you would be likely to end up with ten different devices. As long as they were all working normally under normal conditions, it would be difficult to tell the completed devices apart, or to make deductions about the design of each, because of course they would all behave identically. However, the different devices would be likely to fail in different ways, whether through structural defects or through operation outside normal conditions, and the ways in which they failed would give clues to their design - whether they used analogue or digital technology, whether they used closed loop or open loop control, and so on. If in the future we might wish to claim that our system was similar to that underpinning human consciousness, then it would be prudent to explore not only the similarities when both systems were working correctly, but also the similarities in failure modes when both systems were operating out of specification.

A Key Technical Challenge: Transparency

We have set out our reasons for supposing that models may be useful in achieving intelligence, and that some aspects of operation of certain models may have characteristics resembling those of consciousness. However, we need to find some way of making our systems transparent: unless we can devise some methods of showing that models have been formed, that they are models of certain other entities or processes, and that they are operating in certain ways, we will be unable to give an account of the system's internal workings. In the first experiments described above, it was possible to extract and represent the information in the models in ways that were useful in explaining the system's operation (Figure 2). Maintaining at least this degree of insight into internal processes as our systems grow in complexity is a key technical challenge of our programme.

Is there any general technique that can be employed to detect that a model is present, to show what the model corresponds to, and to show how it is being used? We are not aware of any such general approach, but two recent projects show that it may be possible to obtain such information on a case-by-case basis. In the first, Aleksander and his collaborators present a system including components closely modelled on the visual brain of humans in which the learned internal representation of each visual stimulus in the input array is constrained to form in such a way that it can be displayed in the same terms as the original stimulus for example, the representation of a red square appears as a red square. (Aleksander *et al*., 1999; 2001). After training, such internal representations can be elicited even in the absence of visual stimulation. When the system is presented with an ambiguous situation in which it has to produce a representation as a response, it is possible to see the progress and resolution of the conflict between alternative representations in terms of the visual stimuli from which they were originally formed. In the second, Edelman and his collaborators use the Darwin VII mobile robot to show how the process of neuronal group selection creates essentially arbitrary groups of cells whose co-activation indicates a response to a particular characteristic of visual input (for example, stripes) regardless of position, size, or orientation (Krichmar and Edelman, 2002). It is possible to see these groups being formed through experience, to see them being activated in the presence of the appropriate stimuli, and to appreciate their role in the control of the robot's behaviour. Although these projects form and use the equivalent of categorical representations, a relatively simple kind of model, they show the identification and characterisation of an internal representation used to inform or control decisions; we aim to extend this approach to deal with more dynamic models, and more complex manipulations and exploitations of such models.

One possible approach might be to take advantage of the fact that it is possible to have a complete record of all inputs, outputs, and internal processes for an artificial system, especially if it is a digital system. It might therefore be possible to train a neural network to identify, and perhaps even to characterise, the external and internal events that correspond to the states and state transitions of the various internal models. Presented with a sequence of activations from such an internal model, the neural network would be required to yield the events in the world that the model was representing. There are many reasons why this might prove to be impossible, but the potential rewards are so great that we feel it should at least be attempted.

Acknowledgments

The ARA VQ simulations and experiments were carried out at the California Institute of Technology by Lerone Banks and Nathan Grey.

References

Aleksander, I., Dunmall, B. and Del Frate, V. (1999), 'Neurocomputational models of visualisation: A preliminary report', IWANN (I), 1999, pp. 798-805.

Aleksander, I., Morton, H. and Dunmall, B. (2001), 'Seeing is believing: Depictive neuromodeling of visual awareness', IWANN (/),2001, pp. 765-71.

Arbib, M.A. (1972), The Metaphorical Brain: An Introduction to Cybernetics as Artificial lntelligence and Brain Theory (Chichester: Wiley-1nterscience).

Baars, BJ. (1988), A Cognitive Theory of Consciousness (Cambridge: Cambridge University Press).

Behrmann, M. (2000), 'The mind's eye mapped onto the brain's matter', Trends in Psychological Science, 9 (2), pp. 50-4. Block, N. (1995), 'On a confusion about a function of consciousness', Behavioral and Brain Sciences, 18 (2), pp. 227-87. Brooks, R.A. (1991), 'Intelligence without representation' , Artificial Intelligence, 47, pp. 139-59.

Churchland, P.S. (2002), 'Self-representation in nervous systems', Science, 296, pp. 308-10. Clark, A. and Grush, R. (1999), 'Towards a cognitive robotics', Adaptive Behavior, 7 (1), pp. 5-16.

Craik, KJ.W. (1943), The Nature of Explanation (Cambridge: Cambridge University Press). Crick, Francis (1994), The Astonishing Hypothesis: The Scientific Search for the Soul (New York: Charles Scribner's Sons).

Cruse, H. (1999), 'Feeling our body: The basis of cognition?', Evolution and Cognition, 162 (5), p.2.

Cyberbotics (2003a), http://www.cyberbotics.com Web pages for Webots; Cyberbotics Sarl, Lausanne.

Cyberbotics (2003b), http://www.cyberbotics.com/products/robots/khepera.html Web pages for the Khepera; Cyberbotics Sarl, Lausanne.

Damasio, A.R. (1999), The Feeling of What Happens: Body and Emotion in the Making of Consciousness (London: Harcourt Brace & Co).

Dawkins, R. (1976), The Selfish Gene (Oxford: Oxford University Press).

Dawkins, R. (1998), Unweaving the Rainbow (London: Penguin).

Dennett, D.C. (1995), Darwin's Dangerous idea: Evolution and the Meanings of Life (London: Allen Lane, The Penguin Press).

Dreyfus, HL (1992), What Computers Still Can't Do (Cambridge, MA: MIT Press).

Edelman, G.M. (1987), Neural Darwinism: The Theory of Neuronal Group Selection (New York: Basic Books Inc.).

Gallagher, S. and Shear, J. (ed. 1999), Models of the Self (Exeter: Imprint Academic).

Grush, R. (1997), 'The architecture of representation', Philosophical Psychology, 10, pp. 5-23. Grush, R. (2002), 'An introduction to the main principles of emulation: Motor control, imagery, and perception', Technical Report, Philosophy, UC San Diego.

Hauser, M.D. (2000), Wild Minds: What Animals Really Think (New York: Henry Holt). Hesslow, G. (2002), 'Conscious thought as simulation of behaviour and perception', Trends in Cognitive Sciences, 6, pp. 242-7.

Johnson-Laird, P.N. (1983), Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness (Cambridge: Cambridge University Press; Cambridge, MA: Harvard University Press).

Kelly, I., Holland, 0., Scull, M. and McFarland, D. (1999), 'Artificial autonomy in the natural world: Building a robot predator', Proceedings 5th European Conference on Artificial Life (ECAL'99), pp 289-93 (Berlin: Springer-Verlag).

Krichmar, J.L. and Edelman, G.M. (2002), 'Machine psychology: Autonomous behavior, perceptual categorisation, and conditioning in a brain-based device' , Cerebral Cortex, 12, pp. 818-30.

Linaker, F. and Niklasson, L. (2000a), '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), pp. 199-208 (Cambridge, MA: MIT Press).

Linaker, F. and Niklasson, L. (2000b), 'Time series segmentation using an adaptive resource allocating vector quantization network based on change detection', in Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (UCNN 2000), pp. 323-8 (IEEE Computer Society).

Metzinger, T. (2000), 'The subjectivity of subjective experience: A representationalist analysis of the first person perspective', in The Neural Correlates of Consciousness, ed. T. Metzinger (Cambridge, MA: MIT Press).

Minsky, ML (1968), 'Matter, mind, and models', in Semantic Information Processing, ed. M.L. Minsky (Cambridge, MA: MIT Press).

Nauck, D., Klawonn, F. and Kruse, R. (1997), Foundations of Neuro-Fuzzy Systems (Chichester: Wiley).

Norretranders, T. (1998), The User Illusion: Cutting Consciousness Down To Size, trans. J. Sydenham (London: Allen Lane, Penguin Press).

Panksepp, J. (1998), 'The periconscious substrates of consciousness: Affective states and the evolutionary origins of the self', Journal of Consciousness Studies,S (5-6), pp. 566-82.

Passino, K.M. and Yurkovich, S. (1998), Fuzzy Control (New York: Addison-Wesley).

Perlis, D. (1997), 'Consciousness as self-function', Journal of Consciousness Studies, 4 (5-6), pp. 509-25.

Ramachandran, V.S. and Blakeslee, S. (1998), Phantoms in the Brain: Human Nature and the Architecture of the Mind (London: Fourth Estate).

Richter, W., Somorjai, R., Summers, R., Jarmasz, M., Menon, R.S., Gati, J.S., Georgopoulos, A.P., Tegeler, c., Ugurbil, K. and Kim, S.G. (2000), 'Motor area activity during mental rotation studied by time-resolved single trial fMRI', Journal of Cognitive Neuroscience, 12 (2), pp. 310-20.

Rolls, E.T. (1999), The Brain and Emotion (Oxford: Oxford University Press).

Sporns, 0., Almassy, N. and Edelman, G.M. (2000), 'Plasticity in value systems and its role in adaptive behavior', Adaptive Behavior, 8 (2), pp. 129-48.

Steinkiihler, U., Beyn, W-J. and Cruse, H. (1995), 'A simplified MMC model for the control of an arm with redundant degrees of freedom', Neural Processing Letters, 2, pp. 11-15.

Strawson, G. (1999), 'The self and the SESMET', in Gallagher and Shear (1999).

Tani, J. (1998), 'An interpretation of the "self' from the dynamical systems perspective: A constructivist approach', Journal of Consciousness Studies, S (5-6), pp. 516-42.

von Uexkull, J. (1934/1957), 'A stroll through the worlds of animals and men', in Instinctive Behaviour: The Development of a Modern Concept, ed. C.H. Schiller and K.S. Lashley (New York: International University Press).

Watt, D.F. (2000), 'Emotion and consciousness II: A review of "The Feeling of What Happens"', Journal of Consciousness Studies, 7 (3), pp. 72-84.

Werbos, P.J. (1990), 'A menu of designs for reinforcement learning over time', in Neural Networks for Control, ed. R.S. Sutton, W. T Miller, III and PJ.

Werbos (Cambridge, MA: MIT Press). Werbos, PJ. (1992), 'Approximate dynamic programming for real-time control and neural modeling', in Handbook of Intelligent Control, ed. D. White, A. and D.A. Sofge (Van Nostrand Reinhold).

Wilkinson, S. (2000), 'Gastrobots: Benefits and challenges of microbial fuel cells in food powered robot applications', J. Autonomous Robots, Sept. 2000

Wolpert, D.M. and Ghahramani, Z. (2000), 'Computational principles of movement neuroscience', Nature Neuroscience Supp. vol 3, pp. 1212-17.