

Chapter 4

Brains and Behaviour

She looked liked she learned to dance,
From a series of still pictures.
— Elvis Costello, *Satellite*

4.1 Neuropsychology and Neuroethology

There are two strategies for tackling a really difficult problem. The first is to break it down into discrete sub-problems, solve each of these in isolation, and hope that the partial solutions can be added together to form an explanation of the whole. Many problems are suited to this approach; it underlies our spectacular progress in developing new technology, for example. This success has encouraged its application to many other areas, including the philosophy of mind and cognitive science. Thus the phenomenal complexity of human thought is broken down into the sub-problems of perception, language-use, logical reasoning, concept formation, emotions, motor co-ordination, associative learning, social intelligence, etc, and each of these ‘modules’ are then studied and analysed separately.

The alternative strategy is to start with the simplest possible example of the whole phenomenon, endeavour to understand it as a unified whole, and then consider more and more complex examples, noting qualitative and quantitative changes in behaviour as we do so. There is good reason to believe that cognition is better suited to this type of approach — after all, the strict modularity assumed by cognitive science bears little relation to how brains evolve, develop, learn, or are used in practice. The conclusion is that we should not start by considering isolated competencies of large, cognitively complex, creatures, but rather we should start by considering the whole of simple ones. *Why not the whole iguana?*, as Dennett put it (1978)¹. Iguanas may lack many of our higher cognitive functions. They are probably not even conscious. Nonetheless they can negotiate a complex environment, find food and mates, avoid predators, and so on. They are a simple, complete, example of an intentional system, and so seem like a good starting point — both for our attempts to study natural cognisers, and also to engineer artificial ones. Therefore chapters 4–5 are mostly concerned with only the simplest types of intentional activity of both animals and

¹Darwin was perhaps thinking along the same lines when he claimed that ‘he who understands baboon would do more toward metaphysics than Locke’.

robots, and I only start to consider 'higher' linguistic abilities in chapter 6.

It is also worth remembering that the vast majority of *human* behaviour is similarly basic. Without the ability to navigate and manipulate our environment, humans would not be able to support higher cognitive functions, either individually or socially. The thin layer of conscious, linguistic, reflective icing tops a very thick practical cake:

It is instructive to reflect on the way in which earth-based biological evolution spent its time. Single-cell entities arose out of the primordial soup roughly 3.5 billion years ago. A billion years passed before photosynthetic plants appeared. After almost another billion and a half years, around 550 million years ago, the first fish and vertebrates arrived, and then insects 450 million years ago. Then things started moving fast. Reptiles arrived 370 million years ago, followed by dinosaurs at 330 and mammals at 250 million years ago. The first primates appeared 120 million years ago and the immediate predecessors to the great apes a mere 18 million years ago. Man arrived in roughly his present form 2.5 million years ago. He invented agriculture a mere 19,000 years ago, writing less than 5000 years ago and 'expert' knowledge over the last few hundred years.

This suggests that problem solving, language, expert knowledge and application, and reason, are all pretty simple once the essence of being and reacting are available. That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction. This part of intelligence is where evolution has concentrated its time — it is much harder. (Brooks, 1991)

Traditional philosophy of mind has, like a spoilt child, tried to pick the icing off the cake. It has concentrated on our ability to contemplate the world in isolation from our more fundamental and precious ability to act on and within it. And unfortunately this attitude has been encouraged by the development of powerful brain imaging techniques such as CAT, PET, and especially NMR (Tootell et al., 1995) that can map neuronal activity across the brain while the (usually) human subject remains perfectly still and performs a simple psychological task. The flood of data that these experiments generate is very impressive, but it is still unclear whether it is particularly *useful*. Although correlations between psychological state and brain activity can be demonstrated, the link is not made intelligible. There is no sense of an explanation of *why* or *how* the brain activity produces the psychological phenomena. Nor can these experiments tell us whether the correlation is significant or epiphenomenal. Lesioning experiments and the study of aphasics may be useful in this last respect, but they do not tell us what aspect of the activity of the disrupted tissue was important, nor why.

The missing explanatory link is *activity*. Explaining how brains work requires understanding how brain states play a causal role in behaviour, not just observing correlations. This is the aim of neuroethology. And in order to understand how brain states can play such a role it is necessary to discover their relationship to the rest of the central nervous system of the animal, from sensor to muscle, and *via* feedback from its environment. The fact that the brain has a body is true, obvious, important, and usually ignored. Understanding the bodily and environmental context of the brain can change our picture of what it is doing:

These observations can be summarised using two contrasting musical metaphors. The nervous system is often seen as the conductor of the body, choosing the program

for the players and directing how they play. The results reviewed above suggest a different metaphor: the nervous system is one of a group of players engaged in jazz improvisation, and the final result emerges from the continued give and take between them. In other words, adaptive behaviour is the result of the continuous interaction between the nervous systems, the body and the environment, each of which have rich, complicated, highly structured dynamics. The role of the nervous system is not so much to direct or program behaviour as to shape it and evoke the appropriate patterns of dynamics from the entire coupled system. As a consequence one cannot assign credit for adaptive behaviour to any one piece of this coupled system. (Beer & Chiel, 1997)

If we want an explanation of the sound of an orchestra we have only to look to the conductor and the score that they are following. The characteristics of the individual musicians are relatively unimportant. But if we want an explanation of the performance of jazz ensemble then no player can be ignored. This is not to imply that such an understanding is impossible, but that it cannot be reduced to being the responsibility of a single isolated element. Similarly, if we want an explanation of how the neural mechanisms of a creature subserves its behaviour, then it is not enough to just observe the activity of a single part, but rather we must understand how it is coupled to the rest of the system — i.e. its body and environment.

To take a simple example, the periodic limb movements involved in most forms of animal locomotion are often presumed to be due to internal central pattern generators which propagate centrifugal signals which control the muscles and limbs. But this ignores the role of environmental feedback in generating the overall activity. For example, if you take a lamprey out of water then the change in resistance means that the same stimulation of the muscles produces a completely different wriggle; therefore recording the output from the central pattern generators in its spinal ganglia will give you only part of the picture. Without understanding the properties and role of the water it is impossible to understand how a lamprey swims.

The problem with neuroethology is that it is very hard. Instead of studying isolated parts of the brain of a creature it is necessary — at least in principle — to understand its entire central nervous system, body and environment. This has tended to limit the growth of neuroethology. As Dawkins points out (1995), there is still a large gap between neurobiology and ethology, maintained by the fact that the two disciplines tend to study different animals and ask different questions. Ethologists gravitate towards large intelligent animals — including humans — with interesting and complex behaviours but with poorly understood neurobiology, whereas neurobiologists favour sea slugs and leeches that ethologists find boring. The most interesting work in neuroethology has occurred somewhere in the middle, with creatures that are large enough to display interesting behaviours but are still simple (and disposable) enough for investigation of the entire neural pathway to be possible: prey-catching in frogs and toads, echolocation in bats, auditory source location in owls, and so on.

The difficulties of neuroethology are made even worse by the fact that it is not enough to investigate the *whole* iguana (or bat, frog, or owl), but that they must also be investigated *the right environmental and behavioural context*. For example, over a period of 50 years from 1926 the visual system of the Horseshoe crab became one of the most thoroughly investigated neurophysiological systems in the animal kingdom. However it was not until the late 1970's that it was discovered that the way that the retina reacts to light follows a circadian rhythm, becoming a mil-

lion times more sensitive at night in order to aid mate detection. This crucial functional property of the nervous system had remained undiscovered whilst the visual system had been investigated in *in vitro* isolation as a lab preparation; instead it required taking measurements from a whole live animal in its native conditions of shallow coastal water at night (Barlow et al., 1984)(Barlow et al., 1986).

The need to get the behavioural environment right can stretch the ingenuity of scientists to the limit. Consider the problems of investigating the neuroethology of locust flight. Some progress had been made by taking microelectrode recordings from paralysed insects, but such artificial conditions tends to produce artefactual results. The only solution was to taking recordings from locusts *while they are flying free*, and this required implanting the insects with microelectrodes that would not disrupt their movements and connected to miniature radio transmitters tied to their backs (Kutsch et al., 1993) — a painstaking, intricate, and very frustrating process. Nonetheless, experiments conducted *in* environmental and behavioural *situ* can yield neurological data that it would not be possible to derive, even in principle, from non-situated investigation. A bird in the bush is worth two in the hand, neuroethologically speaking.

Beer, amongst others, argues that the problem of neuroethology is to understand how central nervous systems are coupled to environments *via* bodies. But the situation is actually more complicated than that. If it were simply the case that behaviour is generated by the coupling between a nervous system and an environment then it would be possible, at least in theory, to study the organism in isolation and then try to determine the result if it were put into a particular environment². The more fundamental problem is that this coupling can change the intrinsic properties of the central nervous system itself. The Hodgkin-Huxley model of neuronal activity, which models neurons as discrete ‘units’ with fixed electrical responses, has been very successful. But this success should not make us forget that neurons are living cells whose seemingly intrinsic properties are affected by the metabolism and biochemistry of the entire body and its environment. In some cases there are clearly stereotyped reflex behaviours — such as escape responses — in which the strong evolutionary pressure to favour fast and reliable performance produces dedicated neural structures with very stable and clearly defined properties. However it is now becoming apparent that modulators — hormones, diffuse neurotransmitters, and less obvious agents such as nitric oxide synthases — can alter even the most seemingly fixed and apparent properties of individual neurons (Harris-Warrick & Marder, 1991):

The effects of modulatory substances can be so profound that cells acquire entirely new properties not seen in the absence of the modulator. The effects of modulators covers the range of intrinsic properties, including increased or decreased excitability, the modulation of spike frequency adaption, the enhancement of post-inhibitory rebound, the induction of plateau potentials, and the expression of intrinsic bursting. (Getting, 1989)

These kinds of modulatory processes are usually ignored when constructing artificial neural networks which model biological neural systems using the formalism of systems theory. These models have often been criticised by biologists for being too simplistic. This may be true, but if this

²For a formal analysis and experimental demonstration of how we can do this for an artificial nervous system see (Jakobi, 1997).

were the sole problem then it could be solved by increasing their complexity and accuracy — as has been done in many cases (Lansner & Liljenström, 1994). The more fundamental problem is that virtually all such models assume a fixed neural structure, comprised of units with fixed electrical responses and connections, or ones that change irreversibly through incremental learning³ — and this assumption is rarely true.

For example Soffe (1993) describes how the same set of neurons drive both the swimming and struggling behaviours in *Xenopus* tadpoles. A tadpole that starts by swimming may, depending on its environment, encounter a predator. This sensory stimulation has the effect of modulating the synaptic connections and intrinsic properties of the motor neurons in its spinal cord, with the result that it starts struggling. Note that this is *not* just the effect of new stimuli provoking new responses, but rather involves a reorganisation of the neural system that subserves behaviour. Therefore in order to properly understand these events we first have to understand the neuronal organisation underlying the initial swimming behaviour. We then have to understand how that behaviour, in that particular environment, results in the creature being threatened. Lastly we have to understand how this results in changes at the level of individual neurons as it starts to struggle. There is thus a dialectical causal cycle, from neuroscience to intentional behaviour *and back again*. The properties of nervous systems are an emergent product of behaviour, as much as *vice versa*. So, for example, if we were to study a *Xenopus* embryo in a lab preparation we would not uncover the mechanism responsible for struggling, nor that for swimming, but rather some biochemical mish-mash of the two. The neurological roots of its behaviour would remain a mystery.

Different behaviours produce, and are produced by, different neurological organisations. You cannot study an organism in one context and be sure that even the most intrinsic neural property that you discover will persist in another. In short, if you want to understand how the brain of an animal works, you have to study it in an appropriate behavioural environment. And there is simply no way round this.

4.2 Representation and Explanation

Neuroethology is a dialogue between neuroscience and ethology, born from the conviction that each must be understood in the light of the other. This implies that your neuroscience will depend on your ethology: your understanding of how a neural mechanism subserves behaviour will depend on how you understand that behaviour. For example Hoyle, in his manifesto for neuroethology (1984), assumes a traditional Lorenzian ethology, complete with Fixed Action Patterns, psychohydraulics, displacement acts and releasers etc. Therefore his neuroscientific investigations concern such issues as the neural mechanisms underlying variations in internal drive and motivation.

However, as many of the peer reviews to Hoyle's article point out, there is a lot more to animal behaviour than those aspects considered by Lorenz. Most neuroethology is concerned with behaviour — or rather *aspects* of behaviour — that should properly be classed as intentional; i.e. those in which internal states are attributed to a creature in order to understand how its behaviour is co-ordinated with respect to objects in the environment. The neuroethological problem is then to explain how this co-ordination is subserved by the central nervous system of the agent; and to do

³Though see (Husbands, 1998) for an interesting counterexample

this we must find some neurophysiological property that is capable of explaining how this intentional aspect of the behaviour is achieved. If the behaviour to be explained is defined with respect to a distal object, then the mechanism that produces it must be understood in the same way. The explanans and explananda must share some common vocabulary in order for the connection to be made intelligible, and a common term that relates behaviour and mechanism is *representation*, by which I mean the way that a functional property, process or entity of an agent (the representational vehicle) plays a role in the intentional behaviour of an agent in virtue of information that it carries about the object (the content).

For example, suppose we are trying to understand how rats manage to relocate sources of food in a laboratory arena — which they can do despite the experimenter's attempts to confuse them by moving landmarks, or even flooding the arena and forcing the animal to swim. This ability cannot be explained by simply mapping neuronal connections from sensory stimuli to motor responses, since both the stimuli and responses will change as the experimenter changes the arena. Rather an explanation must reveal how the rat achieves a 'sense of place' by integrating many sources of information, including recognising landmarks and its sense of its own movement. And a vital part of this was the discovery by O'Keefe and Dostrovsky (1971) that certain hippocampal neurons are selectively active as the animal moves between different locations in an environment — so called 'place cells'.

Now a great deal remains unknown about the role of the hippocampus in spatial navigation, despite a huge amount of empirical investigation (see (McNaughton, 1996) and (Knierim et al., 1995)). We must admit that, although we know that there are striking correlations between the animal's perceived location and particular neural activity, we do not know how those correlations fit into the entire sensory-motor system of the rat. For example, one of the most perplexing problems is how the same area of hippocampus can serve as a map for many different arenas simultaneously, depending on other contextual cues. Indeed it is quite possible that once we get the bigger picture we will find that the simple place-cells that we naively thought played a role are just some epiphenomenal by-products of a more complex, higher-level, picture. Nonetheless, unless and until these problems are solved we will not have a proper explanation of how the rat navigates its environment.

Place-cells are an example of how a single neuron, or localised group of neurons, may play a representational role (Barlow, 1972). But there is no reason why this should be the case in general. At the start of the last century Sherrington argued that behaviourally significant aspects of neuronal activity may be organised at a higher level than that of the single neuron (1906). For example Freeman (1985) has demonstrated how oscillations in the vertebrate olfactory bulb involving up to a quarter of a million neurons can encode odorant information. These oscillations have a dominant frequency typically around 40-90Hz, but the refractory period of a typical neuron restricts it to producing action potentials at around 5-10Hz. Therefore the bulbar oscillation must be the result of *co-ordinated activity across the entire bulb*; it cannot be a purely epiphenomenal aggregate effect. For each individual neuron the only thing oscillating at 40-90Hz is the *probability* that it will fire, since it can *actually* only produce an action potential every 10 cycles or so. The large-scale oscillations emerge from the mass action of the whole, but in turn they entrain the activity of the individuals (Faith, 1995).

Odorant information only exists at a level of organisation much higher than that of the single neuron. Indeed there is no reason in principle why a representational vehicle could not be a state or process defined over an entire central nervous system, in the same way that the pressure of a gas is subserved by an aggregate property defined over all its constituent molecules. But at whatever level of organisation we discover them, representations are a necessary term in an explanation of how a neural mechanism produces intentional behaviour. Unless we can understand how the organism represents aspects of its environment we cannot understand *how* its behaviour is coordinated with respect to those aspects, we only know *that* it is. Of course, barring miracles, there must be an explanation of how specific stimuli provoke specific responses. But this does not provide an explanation of the intentional behaviour *per se*; only representations can do this. This issue will be discussed in more philosophical detail in the next chapter, but the same point has also recently taken a more practical form.

4.3 South Coast AI

Dretske once claimed that ‘if you can’t make one, you don’t know how it works’, and theories about how intelligent behaviour is produced have always been tested in the tribunal of construction. So, for example, computationalism as a theory of mind naturally led to computationalism as a way of building artificial intelligences: an interdisciplinary research program that was born at the famous Dartmouth Conference on the East coast of the US in 1956.

The cornerstone of computationalism is that intelligence is necessarily grounded in a formal symbol system or language of thought — a direct descendent of Frege’s insistence that the starting point for a philosophy of mind is the formal study of language. Computationalism therefore implies that the key to building an artificial intelligence is a system that manipulates symbols in the right way, as enshrined in Newell and Simon’s Physical Symbol System Hypothesis (1972). If the computationalist wants to build a robot that can physically interact with the world then the trick is to connect the symbol manipulator to distinct perceptual ‘modules’ that generate symbolic representations of the world which are then manipulated syntactically to produce a set of symbols representing a plan, and this is then transformed into physical movements by the motor modules (Fodor, 1983). The sensory and motor modules are ‘the stupidity in the system’ (Karmiloff-Smith, 1994), while the real intelligence resides in the symbol manipulation. According to this view, the sensory and motor links to the outside world can be eliminated and cognition understood as a purely disembodied phenomenon; hence the inputs and outputs to most AI systems are symbols with no essential connection to the states of the world they are supposed to represent.

However, practical problems within AI raise concomitant doubts about computationalism as a theory of mind. In particular, although AI has been spectacularly successful on tasks (such as playing chess) that humans find very difficult, it had been relatively unsuccessful on tasks (such as simple social language use and sensory-motor co-ordination) that humans find very easy. The first anti-AI wind blew from Berkeley in the West with the publication of the Dreyfus brothers’ *What Computers Cannot Do* (1972). This challenged the fundamental assumptions of Anglo-American analytic philosophy on which computationalist AI was built, and pointed to an alternative philosophical tradition that included the existential phenomenology of Heidegger and Merleau-Ponty and the anti-logicism of the later Wittgenstein. This critique of East Coast AI was soon joined by

others that took ideas from biology (Maturana & Varela, 1980) (Winograd & Flores, 1986), Soviet Psychology's emphasis on activity (Norman, 1993)(Wertsch, 1981), and even concepts taken from Zen Buddhism (Varela, Thompson, & Rosch, 1991).

But Dretske's claim still haunts. The West Coast may provide effective critiques of computationalism, but can it yield a practical guide to building artificial intelligences? For a while it seemed as though connectionism might provide a suitable alternative (Dreyfus & Dreyfus, 1988), but this has (usually) repeated the computationalist assumption that cognition is the transformation of one set of representational symbols into another. The only real difference between this form of connectionism and computationalism is that the former uses a vector algebra, rather than scalar, to manipulate its symbols (Cummins & Schwarz, 1987)(Smolensky, 1988).⁴

Another West-Coast alternative has been to try to understand how orthodox computer systems are embedded and used within a social context (see (Laurel, 1997) for a good example). But this approach does not yield artificially intelligent systems, just ones with better interfaces. A *tamagochi*, for example, may be regarded by its owner as a live sentient creature that deserves care and attention. And such products certainly tell us something interesting about our relationship to 'intelligent' computers. But this hardly constitutes the foundations for a research program into building artificial systems that exhibit the intelligence of animals.

However, if the West Coast is correct to insist that cognitive behaviour cannot be characterised as a formal and abstract input-output mapping then the only way to build a cogniser is to build an agent that physically interacts with its world. Thus there has been a rapid increase in research in robotics that eschews conventional computationalist techniques, variously known as artificial life, behaviour-based robotics, the simulation of adaptive behaviour, *nouvelle AI*, post-modern robotics and so on. However I prefer the term 'South Coast AI', referring to the Artificial Life group of the University of Sussex on the south coast of England. This is not a question of academic priority but rather a recognition of the unusual synthesis of philosophical debate, robot engineering, and neuroethology in this institution, as noted by Keeley (1998).

It has to be said that progress in South Coast AI has been painfully slow compared to that of the East Coast. Computer chess players can beat human grandmasters, but the robot footballers that are a feature of most robotics conferences would scarce trouble a two year old child, let alone Pele. Sometimes it is difficult to see any advance over the work of the pioneers of cybernetics in the 1950's, such as Grey Walter and Ross Ashby (1952), or the robotic thought experiments of the Swiss neuroscientist Valentino Braitenberg (1984), despite many billion-fold increases in computer power now used. A critical observer would be justified in thinking that South Coast AI is on a slow road to nowhere, and that this should tell us something about the theory on which it is based. But the fact is that we simply do not have a good theory to replace computationalism as a guide to constructing intelligent agents. And in the absence of a convincing theory, a thousand robotic flowers have bloomed. South Coast AI at the moment is characterised by a large number of often very small research groups, each working on their own particular techniques with very little sense of constructive, cohesive progress. The only notable exception is MIT's *Cog* project, in which a diverse set of particular solutions to partial problems — such as saccading eyes, reaching

⁴Note that this is a criticism of connectionism considered as a method of mapping one set of representations onto another, rather than the use of artificial neural network to control embodied agents — see below.

for an object, and tensing an arm — are being progressively added to a single humanoid robot, in the hope that humanoid intelligence will one day collectively emerge.

However two of these flowers are of theoretical interest. The first is to copy — or at least take inspiration from — nature, and use the findings of neuroethology to model simple natural sensory-motor systems in robots. This is generally known as computational neuroethology (see (Beer, 1990) and (Cliff, 1991)). The second approach is to artificially evolve neural network controllers for robots using genetic algorithms. This is evolutionary robotics (see (Beer & Gallagher, 1992) and (Harvey et al., 1997)). What both these approaches have in common is that, in the absence of a good theory, they avoid designing robot control systems by hand, and instead leave the design process up to natural, or artificial, selection. The use of representations is thus no longer an *a priori* assumption about how to build intentional agents, but is rather an open empirical question about how they work. And a significant minority of researchers have concluded they are simply not necessary, most notably in Brooks' landmark paper *Intelligence Without Representation* (1991). (Also see (Beer, 1995b), (Harvey, 1992), (Cliff & Noble, 1997), (Van Gelder, 1992) and (Wheeler, 1994) for variants on the same theme.)

However all these objections assume that representations must fit the East Coast model, in which the mechanism of the agent can be neatly carved up into humuncular modules which 'communicate' using a vocabulary of symbolic representations. These modules are fixed, disjoint, and completely general purpose (in the sense that there is a single modular organisation capable of producing all behaviours). For example Wheeler, in discussing the analysis of evolved robot control systems, cites Beer's remark that 'highly distributed and richly interconnected systems [such as evolved neural networks] . . . do not admit of any straightforward functional decomposition into representations and modules which algorithmically manipulate them' (Beer, 1995a, p128) (cited in (Wheeler, 1998)).

However Beer *et al* are shooting at the wrong target. They are correct that such evolved networks do not show a modular decomposition obeying algorithmic rules, and such empirical evidence is a powerful weapon against computational and cognitivist assumptions about the mind. Moreover, central nervous systems are the most complex, non-linear, and feedback-ridden systems we know of and understanding them is rarely 'straightforward', especially when we are trying to understand their interactions with a messy real world environment. (I once asked an ethologist who had studied navigation in insects for many years why he did not encourage students to investigate the neural mechanisms underlying the behaviour he was so interested in. His reply was not that this would be impossible, but that someone could easily spend 20 years on this research and still not get anywhere — and this is for a relatively well understood behaviour in a 'simple' insect.)

However in the last section I emphasised that representations are only defined with respect to, and in the context of, the behaviour of a whole agent within an environment. This implies that (1) there need be no general-purpose algorithmic or representational organisation underlying different behaviours, and (2) that any representational functional organisation is an emergent product of the interaction between an agent and its environment. For example as we saw in the case of the Horseshoe crab and the rat hippocampus, the mode of organisation and 'intrinsic' properties of a neural system may change radically from one behavioural context to another, and there is no reason why representational correlations found in one situation should play a role, or even exist,

in another. If this is taken into account then Brooks' *et al* objections lose their force and we can instead appreciate how the examples of robot control systems, and animal nervous systems, that are often held up as paradigm cases of non-representational intentionality do, in fact, have an emergent representational character (Faith, 1997).

One much-cited example is an experiment conducted at the University of Sussex in which artificial evolution was used to generate not only the control system for a robot, but also a suitable body for it to control (Harvey, Husbands, & Cliff, 1994). The 'fitness' of the robot was judged by its ability approach a white triangular target, whilst avoiding a rectangular one. The successful robot used just two sensors, one with a visual field above the other, and located the triangle by rotating on the spot until just the lower sensor saw white and moving straight ahead. This has the effect of fixating the robot on the oblique edge of the triangle. As the triangle looms up such that both sensors go high, or if the motion causes the edge to be lost, then the robot will start to rotate until the edge can be fixated again. The rotate/move-straight distinction is effected by a single unit that takes an inhibitory connection from the upper sensor and an excitatory link from the lower, and is thus only fully activated when the robot is facing towards the triangle's edge.

Two points about this robot must be noted. The first is that its success depends on having a sensor morphology that is perfectly suited to the targets in its environment. If those targets were shaped even slightly differently then there would be no simple way of using the same eyes to do the same discrimination. The control system is also finely tuned to the types of motor and the timing of rotation: if even just the amount of noise in the system is changed then the whole robot has a tendency to overshoot and end up literally going in circles. Therefore, you cannot understand the brain of the robot without also understanding its body and environment. Nonetheless a crucial part of understanding how it does this is to note the correlation between the triangular target being straight ahead, the activation of a particular unit, and the robot moving straight — a representation, in the sense defined above.

To take another example, Floreano and Mondada describe the artificial evolution of a neural network controller for a robot whose task is to explore a simple arena, returning to a recharging base that is demarcated by a black floor patch and directed by a bright light. It was found that the fittest individual used a hidden node of the network whose activation corresponded to the distance from the base, reaching a maximum when it was 'home'. As the experimenters note:

In this experience the robot autonomously evolved the ability to use the raw sensor data and built an internal representation of the world in order to find the recharging area and return to this place at a given time. This behaviour is based on an accurate evaluation of the battery residual time and on an internal representation of the environment. In fact some of the hidden nodes displayed activation levels that clearly mapped the environment geometry. (Mondada & Floreano, 1996)

Evolutionary robotics is in its infancy, and the tasks it tackles are so simple that in many cases they can be solved by agents with only the most direct stimulus-response reflexes. Indeed the evolutionary strategy is brilliant at finding ingenious stimulus-response solutions to tasks that a human designer would normally insist could only be achieved by forming and manipulating representations of the robot's environment. (This specific question is investigated empirically by (Miglino et al., 1998).) However, as tasks become more complex the use of internal states

that carry information about the environment becomes less and less avoidable (Kirsh, 1991), and even in the very simple cases mentioned above we find that individual units act as very simple representations in mediating interactions between the robot and its world. Indeed Brooks himself later conceded that he was not arguing against representations *per se*, but that he was merely advocating different *types* of representation:

My earlier paper (1991) is often criticised for advocating absolutely no representation of the world within a behaviour-based robot. This criticism is invalid. I make it clear in the paper that I reject traditional Artificial Intelligence representation schemes. I also made it clear that I reject explicit representations of goals within the machine.

There can, however, be representations which are partial models of the world — in fact I mentioned that “individual layers extract only those *aspects* of the world which they find relevant — projections of a representation into a simple subspace”. The form these representations take, within the context of the computational model we are using, will depend on the particular task those representations are to be used for. (Brooks, 1995)

The same softening of anti-representationalist attitudes amongst South Coast engineers can be seen amongst Clark and Wheeler (1998), Scheier and Pfeifer (1998), Bickhard (1998), Calabretta *et al* (1998), and Tani *et al* (1998), who all agree that even very simple intentional behaviours are mediated by representations, but representations that can only be understood in the context of activity. At least one neuroethologist draws a similar lesson, but again confuses rejection of computationalist symbols and modules, with rejection of representations *per se*:

Of course the cognitive approach — the representational paradigm — is a level of interpretation in its own right. At best, it is like Ptolemy’s system of epicycles, which could describe the movements of the planets in sufficient detail; but as we now know, the heliocentric view of the world provides a simpler way of understanding this movement and one that comes closer to what is actually the case. By analogy, the cognitive-map approach might obscure some of the most important computational strategies used by the brain. In general, the brain has evolved not to reconstruct a full representation of the three-dimensional world, but to find particular solutions to particular problems within that world. (Wehner, Michel, & Antonsen, 1996)

The representations advocated by both Wehner and Brooks are not general-purpose symbols syntactically manipulated according to an East Coast algorithm, but rather describe how the sensory-motor transformations required for particular behaviours are achieved. The representational organisation underlying, and emergent within, one behaviour need bear no relation to that underlying another.

In Brooks’ experiments this separation is embodied in a ‘subsumption’ robotic architecture — as used on *Cog* — in which the mechanism is split into largely independent ‘layers’, each of which is connected to both sensors and motors. Evolution, both natural and artificial, does not tend to produce such extreme disjointedness but rather produces mixed bags of tricks made up of particular solutions to particular problems, in which evolved circuitry is used and adapted to new purposes. In either case, in order to understand how these systems achieve robust co-ordination with objects in their environment it is necessary to understand how information about those objects play a role in controlling that behaviour.

It is interesting to note that the most doctrinaire anti-representationalists have been computational neuroethologists and evolutionary roboticists, rather than the biologists who study natural sensory-motor systems. It seems that this position stems from a healthy desire amongst computer scientists to disassociate themselves from the tradition of computationalist artificial intelligence and its Cartesian understanding of representation. Biologists have rarely been tarred with the Cartesian computationalist brush — after all, no-one can accuse them of studying disembodied intelligence — and so seem more comfortable with describing the neural mechanisms that they discover in representational terms (Roitblat, 1994).

Lying behind South Coast anti-representationalism there often lurks the intuition that something is only a representation to the extent that it is part of a generalised symbol system. Without such a system it is assumed that an internal state does not have well-defined semantic properties. They therefore share the cognitivist assumption that representations — and intentionality — are to do with computation, rather than the ability of an agent to actively engage with its world. The philosophical roots, and implications, of this argument will be discussed in chapter 6.

The lesson of South Coast AI is that if you want to build a cogniser, you shouldn't start by making up fancy data-structures, since without a body they are both meaningless and useless. Moreover, just bolting on sensory and motor modules will rarely succeed in effectively tying a symbol system to the world, since those contents that a human intuition assigns to the symbols are unlikely to be the ones that its crude body can make available. This was the problem of East Coast robotics, as exemplified by *Shakey* (Nilsson, 1984).

Shakey was a mobile robot that could move blocks round a set of rooms, according to typed instructions. At its heart was a predicate calculus model of its environment, manipulated by a means-end problem solver (Newell & Simon, 1972), and generated from a camera image of its environment — ‘a series of still pictures’, in Costello's phrase. However the problem of producing a symbolic representation of its environment meant that the rooms had to be specially designed to be as visually simple as possible, with flat floors, evenly coloured surfaces, careful lighting, few obstructions, and so on. Although *Shakey* worked, it proved impossible to generalise its success to more realistic environments. The moral is to start by getting the body right, and concentrate on tying it to the world; representations will be the emergent result, as the evolutionary roboticists have repeatedly found. The East Coast approach to artificial intelligence is like noting that a good Emmenthal cheese invariably has holes in it, and concluding that the starting point for making one is to glue pockets of air together. The South Coast approach is to start with the cheese. If you get this right, then you find you get the holes for free.

The limiting factor in our development of intelligent artificial creatures is not the computational power of their ‘brains’, but in the more basic engineering technology of their bodies. Current robot engineers use roughly the same motor and sensor technology that the pioneers of cybernetics did, and yet this is where the real problems of embodied intelligence lie. Therefore we should not be surprised at the slow progress. Rod Brooks draws an illuminating comparison between the development of computer technology, and that of jet airliners. The power, capacity, and speed of the former have doubled roughly every 18 months, whereas the same improvement in the latter has taken almost 40 years. We should expect the development of South Coast AI to be more like that of airliners than computers, and for similar reasons. Building successful robots

depends more on the ‘hard’ engineering of bodies than on the ‘soft’ engineering of brains.

4.4 Conclusion

In order to understand how neural mechanisms can underlie intentional behaviour it is necessary to understand how they can carry information about the environment of the organism. This requires that we investigate the entire causal loop, involving brains, bodies and environment. Moreover, this system must be investigated *in vivo*, since the relevant neurological properties may only exist in the appropriate behavioural context. The same lesson applies when constructing artificial intentional systems: we cannot start from an isolated representational module that approximates humanoid problem solving, since without a humanoid body it will have no effective connection to the world that it is supposed to represent.