Robotics: Philosophy of Mind using a Screwdriver

Inman Harvey
School of Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QH, UK

Abstract

The design of autonomous robots has an intimate relationship with the study of autonomous animals and humans — robots provide a convenient puppet show for illustrating current myths about cognition. Like it or not, any approach to the design of autonomous robots is underpinned by some philosophical position in the designer. Whereas a philosophical position normally has to survive in debate, in a project of building situated robots one's philosophical position affects design decisions and is then tested in the real world — "doing philosophy of mind with a screwdriver".

Traditional Good Old Fashioned Artificial Intelligence (GOFAI) approaches have been based on what is commonly called a Cartesian split between body and mind — though the division goes back at least to Plato. The Dynamical Systems approach to cognition, and to robot design, draws on other philosophical paradigms. We shall discuss how such varied philosophers as Heidegger, Merleau-Ponty or Wittgenstein, in the improbable event of them wanting to build robots, might be tempted to set about the task.

1 Introduction

Car manufacturers need robots that reliably and mindlessly repeat sequences of actions in some well-organised environment. For many other purposes autonomous robots are needed that will behave appropriately in a disorganised environment, that will react adaptively when faced with circumstances that they have never faced before. Planetary exploring robots, such as the Sojourner robot sent to Mars, cannot afford to wait the long time needed for

radio communication with people on earth for consultation on every individual move they make. The user of a semi-autonomous wheelchair should be able to delegate the same sort of decisions that a horse rider delegates to her horse — how to manoeuvre around obstacles and react instinctively to other traffic. We want such robots to behave to some extent intelligently, or adaptively — in fact to behave in some small part as if they had a mind of their own.

It has been tempting to think of this as 'merely' a technical, scientific problem - we should study in an objective, scientific fashion the basic requirements for adaptive intelligence, and then systematically engineer into our robots what we have found to be necessary. But like it or not, any approach to the understanding of cognition and adaptive intelligence, and hence to the design of autonomous robots, is inevitably framed within some philosophical position in the scientist or designer. In a project of building situated robots one's philosophical position affects design decisions and is then tested in the real world — "doing philosophy of mind with a screwdriver".

We use basic working metaphors to make sense of scientific theories; billiard balls and waves on a pond have been much used in physics. The metaphor of animals or even humans as machines, as comparable to the technical artefacts that we construct, is a powerful one. When we try to build autonomous robots they are almost literally puppets acting to illustrate our current myths about cognition. The word 'myth' sounds possibly derogatory as it often implies a fiction or half-truth; it is not intended as such here. I am merely trying to emphasise that our view of cognition is a human-centred view, from the end of the second millennium, some 4 billion years after the origin of life on this planet.

Someone coming from the conventional scientific perspective may suggest that our current cultural context is irrelevant. After all, objectivity in science is all to do with discounting the accidental perspectives of an observer and discovering universal facts and laws that all observers can agree on, from whatever place and time in which they are situated. However, to pursue this line too far leads one into a paradox. What theories (if any) did the organisms of 2 billion years ago have about the cognitive abilities of their contemporaries? Clearly nothing like ours. What theories might our descendants (if any) 2 billion years hence have about the cognition of *their* contemporaries? It would be arrogant, indeed unscientific, to assume that they would be similar to ours.

The Copernican revolutions in science have increased the scope of our objective understanding of the world by recognising that our observations are not context-free. Copernicus and Galileo used their imagination, and

speculated what the solar system might look like if our view from the planet Earth was not a privileged view from the fixed centre of the universe, but merely one possible perspective amongst many. Darwin opened up the way to appreciating that Homo Sapiens is just one species amongst many, with a common mode of evolutionary development from one common origin. Einstein brought about a fresh Copernican revolution with Special Relativity, showing how our understanding is increased when we abandon the idea of some unique fixed frame of reference for measuring the speed of any object. The history of science shows a number of advances, now generally accepted, stemming from a relativist perspective that (surprisingly) is associated with an objective stance toward our role as observers.

Cognitive science seems one of the last bastions to hold out against a Copernican, relativist revolution. In this paper I will broadly distinguish between the pre-Copernican views associated with the computationalist approach of classical Good Old Fashioned Artificial Intelligence (GOFAI), and the contextual, situated approaches of nouvelle AI. The two sides will be rather crudely portrayed, with little attempt to distinguish the many differing factions that can be grouped under one flag or the other. The different philosophical views will be associated with the direct implications that they have for the design of robots. It is worth mentioning that Brooks' recent collection of his early papers on robotics (Brooks, 1999) explicitly divides the eight papers into four under the heading of 'Technology' and four as 'Philosophy' - though the division is somewhat arbitrary as the two aspects go together throughout.

2 Cartesian or Classical approaches to Robotics

Descartes, working in the first half of the seventeenth century, is considered by many to be the first modern philosopher. A scientist and mathematician as much as philosopher, his ideas laid the groundwork for much of the way we view science today. In cognitive science the term 'Cartesian' has, perhaps rather unfairly to Descartes, come to exclusively characterise a set of views that treat the division between the mental and the physical as fundamental — the Cartesian cut (Lemmen, 1998). One form of the Cartesian cut is the dualist idea that these are two completely separate substances, the mental and the physical, which can exist independently of each other. Descartes proposed that these two worlds interacted in just one place in humans, the pineal gland in the brain. Nowadays this dualism is not very respectable, yet the common scientific assumption rests on a variant of this Cartesian cut: that the physical world can be considered completely

objectively, independent of all observers.

This is a different kind of objectivity from that of the Copernican scientific revolutions mentioned above. Those relied on the absence of any privileged position, on intersubjective agreement between observers, independent of any specific observer. The Cartesian objectivity assumes that there just is a way the world is, independent of any observer at all. The scientist's job, then, is to be a spectator from outside the world, with a God's-eye view from above.

When building robots, this leads to the classical approach where the robot is also a little scientist-spectator, seeking information (from outside) about how the world is, what objects are in which place. The robot takes in information, through its sensors; turns this into some internal representation or model, with which it can reason and plan; and on the basis of this formulates some action that is delivered through the motors. Brooks calls this the SMPA, or sense-model-plan-act architecture (Brooks, 1999).

The 'brain' or 'nervous system' of the robot can be considered as a Black Box connected to sensors and actuators, such that the behaviour of the machine plus brain within its environment can be seen to be intelligent. The question then is, 'What to put in the Black Box?' The classical computationalist view is that it should be computing appropriate outputs from its inputs. Or possibly they may say that whatever it is doing should be *interpretable* as doing such a computation.

The astronomer, and her computer, perform computational algorithms in order to predict the next eclipse of the moon; the sun, moon and earth do not carry out such procedures as they drift through space. The cook follows the algorithm (recipe) for mixing a cake, but the ingredients do not do so as they rise in the oven. Likewise if I was capable of writing a computer program which predicted the actions of a small creature, this does not mean that the creature itself, or its neurons or its brain, was consulting some equivalent program in 'deciding what to do'.

Formal computations are to do with solving problems such as 'when is the eclipse?'. But this is an astronomer's problem, not a problem that the solar system faces and has to solve. Likewise, predicting the next movement of a creature is an animal behaviourist's problem, not one that the creature faces. However, the rise of computer power in solving problems naturally, though regrettably, led AI to the view that cognition equalled the solving of problems, the calculation of appropriate outputs for a given set of inputs. The brain, on this view, was surely some kind of computer. What was the problem that the neural program had to solve? — the inputs must be sensory, but what were the outputs?

Whereas a roboticist would talk in terms of motor outputs, the more

cerebral academics of the infant AI community tended to think of plans, or representations, as the proper outputs to study. They treated the brain as the manager who does not get his own hands dirty, but rather issues commands based on high-level analysis and calculated strategy. The manager sits in his command post receiving a multitude of possibly garbled messages from a myriad sensors and tries to work out what is going on. Proponents of this view tend not to admit explicitly, indeed they often deny vehemently that they think in terms of a homunculus in some inner chamber of the brain, but they have inherited a Cartesian split between mind and brain and in the final analysis they rely on such a metaphor.

3 What is the Computer Metaphor?

The concepts of computers and computations, and programs, have a variety of meanings that shade into each other. On the one hand a computer is a formal system with the same powers as a Turing Machine (...assuming the memory is of adequate size). On the other hand a computer is this object sitting in front of me now, with screen and keyboard and indefinite quantities of software.

A program for the formal computer is equivalent to the pre-specified marks on the Turing machine's tape. For a given starting state of this machine, the course of the computation is wholly determined by the program and the Turing machine's transition table; it will continue until it halts with the correct answer, unless perhaps it continues forever — usually considered a *bad thing*!

On the machine on my desk I can write a program to calculate a succession of co-ordinates for the parabola of a cricket-ball thrown into the air, and display these both as a list of figures and as a curve drawn on the screen. Here I am using the machine as a convenient fairly user-friendly Turing machine.

However most programs for the machine on my desk are very different. At the moment it is (amongst many other things) running an editor or word-processing program. It sits there and waits, sometimes for very long periods indeed, until I hit a key on the keyboard, when it virtually immediately pops a symbol into an appropriate place on the screen; unless particular control keys are pressed, causing the file to be written, or edits to be made. Virtually all of the time the program is waiting for input, which it then processes near-instantaneously. In general it is a *good thing* for such a program to continue for ever, or at least until the exit command is keyed in.

The cognitivist approach asserts that something with the power of a

Turing machine is both necessary and sufficient to produce intelligence; both human intelligence and equivalent machine intelligence. Although not usually made clear, it would seem that something close to the model of a word-processing program is usually intended; i.e., a program that constantly awaits inputs, and then near-instantaneously calculates an appropriate output before settling down to await the next input. Life, so I understand the computationalists to hold, is a sequence of such individual events, perhaps processed in parallel.

4 Time in Computations and in Connectionism

One particular aspect of a computational model of the mind which derives from the underlying Cartesian assumptions common to traditional AI is the way in which the issue of *time* is swept under the carpet — only the sequential aspect of time is normally considered. In a standard computer operations are done serially, and the lengths of time taken for each program step are for formal purposes irrelevant. In practice for the machine on my desk it is necessary that the time-steps are fast enough for me not to get bored waiting. Hence for a serial computer the only requirement is that individual steps take as short a time as possible. In an ideal world any given program would be practically instantaneous in running, except of course for those unfortunate cases when it gets into an infinite loop.

The common connectionist assumption is that a connectionist network is in some sense a parallel computer. Hence the time taken for individual processes within the network should presumably be as short as possible. They cannot be considered as being effectively instantaneous because of the necessity of keeping parallel computations in step. The standard assumptions made fall into two classes.

1. The timelag for activations to pass from any one node to another it is connected to, including the time taken for the outputs from a node to be derived from its inputs, is in all cases exactly one unit of time (e.g. a back-propagation, or an Elman network).

2. Alternatively, just one node at a time is updated independently of the others, and the choice of which node is dealt with next is stochastic (e.g. a Hopfield net or a Boltzmann machine).

The first method follows naturally from the computational metaphor, from the assumption that a computational process is being done in parallel.

The second method is closer to a dynamical systems metaphor, yet still computational language is used. It is suggested that a network, after appropriate training, will when presented with a particular set of inputs then sink into the appropriate basin of attraction which appropriately classifies them. The network is used as either a distributed content-addressable memory, or as a classifying engine, as a module taking part in some larger-scale computation. The stochastic method of relaxation of the network may be used, but the dynamics of the network are thereby made relatively simple, and not directly relevant to the wider computation. It is only the stable attractors of the network that are used. It is no coincidence that the attractors of such a stochastic network are immensely easier to analyse than any non-stochastic dynamics.

It might be argued that connectionists are inevitably abstracting from real neural networks, and inevitably simplifying. In due course, so this argument goes, they will slowly extend the range of their models to include new dimensions, such as that of time. What is so special about time — why cannot it wait? Well, the simplicity at the formal level of connectionist architectures which need synchronous updates of neurons disguises the enormous complexity of the physical machinery needed to maintain a universal clock-tick over distributed nodes in a physically instantiated network. From the perspective advocated here, clocked networks form a particular complex subset of all real-time dynamical networks ones need be, and if anything *they* are the ones that should be left for later (van Gelder, 1992).

A much broader class of networks is that where the timelags on individual links between nodes is a real number which may be fixed or may vary in a similar fashion to weightings on such links[1]. A pioneering attempt at a theory that incorporates such timelags as an integral part is given in (Malsburg and Bienenstock, 1986).

In neurobiological studies the assumption seems to be widespread that neurons are passing information between each other 'encoded' in the rate of firing. By this means it would seem that real numbers could be passed, even though signals passing along axons seem to be all-or-none spikes. This assumption is very useful, indeed perhaps invaluable, in certain areas such as early sensory processing. Yet it is perverse to assume that this is true throughout the brain, a perversity which while perhaps not caused by the computational metaphor is certainly aided by it. Experiments demonstrating that the individual timing of neuronal events in the brain, and the temporal coincidence of signals passing down separate 'synfire chains', can

[1]For a simple model without loss of generality any time taken for outputs to be derived from inputs within a node can be set to zero, by passing any non-zero value on instead to the links connected to that node.

be of critical importance, are discussed in (Abeles, 1982).

5 What is a Representation?

The concept of symbolic reference, or representation, lies at the heart of analytic philosophy and of computer science. The underlying assumption of many is that a real world exists independently of any given observer; and that symbols are entities that can 'stand for' objects in this real world — in some abstract and absolute sense. In practice, the role of the observer in the act of representing something is ignored.

Of course this works perfectly well in worlds where there is common agreement amongst all observers — explicit or implicit agreement — on the usages and definitions of the symbols, and the properties of the world that they represent. In the worlds of mathematics, or formal systems, this is the case, and this is reflected in the anonymity of tone, and use of the passive tense, in mathematics. Yet the dependency on such agreement is so easily forgotten — or perhaps ignored in the assumption that mathematics is the language of God.

A symbol P is used by a person Q to represent, or refer to, an object R to a person S. Nothing can be referred to without somebody to do the referring. Normally Q and S are members of a community that have come to agree on their symbolic usages, and training as a mathematician involves learning the practices of such a community. The vocabulary of symbols can be extended by defining them in terms of already-recognised symbols.

The English language, and the French language, are systems of symbols used by people of different language communities for communicating about their worlds, with their similarities and their different nuances and clichés. The languages themselves have developed over thousands of years, and the induction of each child into the use of its native language occupies a major slice of its early years. The fact that, nearly all the time we are talking English, we are doing so to an English-speaker (including when we talk to ourselves), makes it usually an unnecessary platitude to explicitly draw attention to the community that speaker and hearer belong to.

Since symbols and representation stand firmly in the linguistic domain, another attribute they possess is that of arbitrariness (from the perspective of an observer external to the communicators). When I raise my forefinger with its back to you, and repeatedly bend the tip towards me, the chances are that you will interpret this as 'come here'. This particular European and American sign is just as arbitrary as the Turkish equivalent of placing the hand horizontally facing down, and flapping it downwards. Different actions

or entities can represent the same meaning to different communities; and the same action or entity can represent different things to different communities. In Mao Tse-Tung's China a red traffic light meant *GO*.

In the more general case, and particularly in the field of connectionism and cognitive science, when talking of representation it is imperative to make clear who the users of the representation are; and it should be possible to at a minimum suggest how the convention underlying the representation arose. In particular it should be noted that where one and the same entity can represent different things to different observers, conceptual confusion can easily arise. When in doubt, always make explicit the Q and S when P is used by Q to represent R to S.

In a computer program a variable `pop_size` may be used by the programmer to represent (to herself and to any other users of the program) the size of a population. Inside the program a variable i may be used to represent a counter or internal variable in many contexts. In each of these contexts a metaphor used by the programmer is that of the program describing the actions of various homunculi, some of them keeping count of iterations, some of them keeping track of variables, and it is within the context of particular groups of such homunculi that the symbols are representing. But how is this notion extended to computation in connectionist networks?

6 Representation in Connectionism

When a connectionist network is being used to do a computation, in most cases there will be input, hidden and output nodes. The activations on the input and output nodes are decreed by the connectionist to represent particular entities that have meaning for her, in the same way as `pop_size` is in a conventional program. But then the question is raised — 'what about internal representations?'.

If a connectionist network is providing the nervous system for a robot, a different interpretation might be put on the inputs and outputs. But for the purpose of this section, the issues of internal representation are the same.

All too often the hidden agenda is based on a Platonic notion of representation — what do activations or patterns of activations represent in some absolute sense to God? The behaviour of the innards of a trained network are analysed with the same eagerness that a sacrificed chicken's innards are interpreted as representing ones future fate. There is however a more principled way of talking in terms of internal representations in a network, but a way that is critically dependent on the observer's decomposition of that

network. Namely, the network must be decomposed by the observer into two or more modules that are considered to be communicating with each other by means of these representations.

Where a network is explicitly designed as a composition of various modules to do various subtasks (for instance a module could be a layer, or a group of laterally connected nodes within a layer), then an individual activation, or a distributed group of activations, can be deemed to represent an internal variable in the same way that i did within a computer program. However, unlike a program which wears its origins on its sleeve (in the form of a program listing), a connectionist network is usually deemed to be internally 'nothing more than' a collection of nodes, directed arcs, activations, weights and update rules. Hence there will usually be a large number of possible ways to decompose such a network, with little to choose between them; and it depends on just where the boundaries are drawn just who is representing what to whom.

It might be argued that some ways of decomposing are more 'natural' than others; a possible criterion being that two sections of a network should have a lot of internal connections, but a limited number of connecting arcs between the sections. Yet as a matter of interest this does not usually hold for what is perhaps the most common form of decomposition, into layers. The notion of a distributed representation usually refers to a representation being carried in parallel in the communication from one layer to the next, where the layers as a whole can be considered as the Q and S in the formula "P is used by Q to represent R to S".

An internal representation, according to this view, only makes sense relative to a particular decomposition of a network chosen by an observer. To assert of a network that it contains internal representations can then only be justified as a rather too terse shorthand for asserting that the speaker proposes some such decomposition. Regrettably this does not seem to be the normal usage of the word in cognitive science, yet I am not aware of any well-defined alternative definition.

7 Are Representations Needed?

With this approach to the representation issue, then any network can be decomposed (in a variety of ways) into separate modules that the observer considers as communicating with each other. The interactions between such modules can *ipso facto* be deemed to be mediated by a representation. Whether it is useful to do so is another matter.

Associated with the metaphor of the mind (or brain, or an intelligent

machine) as a computer go assumptions of functional decomposition. Since a computer formally manipulates symbols, yet it is light waves that impinge on the retina or the camera, surely (so the story goes) some intermediate agency must do the necessary translating. Hence the traditional decomposition of a cognitive system into a perception module, which takes sensory inputs and produces a world model; this is passed onto a central planning module which reasons on the basis of this world model; passing on its decisions to an action module which translates them into the necessary motor actions. This functional decomposition has been challenged, and an alternative behavioural decomposition proposed, by Brooks in, e.g., (Brooks, 1999).

In particular, the computationalist or cognitivist approach seems to imply that communication between any such modules is a one-way process; any feedback loops are within a module. Within for instance back-propagation, the backward propagation of errors to adjust weights during the learning process is treated separately from the forward pass of activations. This helps to maintain the computational fiction, by conceptually separating the two directions, and retaining a feed-forward network. But consider the fact that within the primate visual processing system, when visualised as a network, there are many more fibres coming 'back' from the visual cortex into the Lateral Geniculate Nucleus (LGN) than there are fibres going from the retina to the LGN in the 'correct' direction. How does the computationalist make sense of this?

Marr (in (Marr, 1977), reprinted in (Boden, 1990)) classifies AI theories into Type 1 and Type 2, where a Type 2 theory can only solve a problem by the simultaneous action of a considerable number of processes, *whose interaction is its own simplest description.* It would seem that type 2 systems can only be decomposed arbitrarily, and hence the notion of representation is less likely to be useful. This is in contrast to a Type 1 theory, where a problem can be decomposed into a form that an algorithm can be formulated to solve, by *divide and conquer.* Type 1 theories are of course the more desirable ones when they can be found, but it is an empirical matter whether they exist or not. In mathematics the 4-colour theorem has been solved in a fashion that requires a large number of special cases to be exhaustively worked out in thousands of hours of computation (Appel and Haken, 1989). It is hoped that there were no hardware faults during the proof procedure, and there is no way that the proof as a whole can be visualised and assessed by a human. There is no *a priori* reason why the workings of at least parts of the brain should not be comparably complex, or even more so[2]. This

[2]For the purposes of making an intelligent machine or robot, it has in the past seemed obvious that only Type 1 techniques could be proposed. However evolutionary techniques

can be interpreted as: there is no *a priori* reason why all parts of the brain should be in such a modular form that representation-talk is relevant. The answer to the question posed in the title of this section is *no*. This does not rule out the possibility that in some circumstances representation-talk *might* be useful, but it is an experimental matter to determine this.

8 Alternatives to Cartesianism

Hubert Dreyfus came out with a trenchant criticism of the classical AI computationalist approach to cognition in the 1960s. He came from a very different set of philosophical traditions, looking for inspiration to twentieth century philosophers such as Heidegger, Merleau-Ponty and Wittgenstein. Initially he produced a report for the RAND Corporation provocatively entitled 'Alchemy and Artificial Intelligence' in 1965 (Dreyfus, 1965). Later, more popular, books are (Dreyfus, 1972) and with his brother (Dreyfus and Dreyfus, 1986). These are amongst the easiest ways for somebody with a conventional background in computer science, cognitive science or robotics to approach the alternative set of philosophical views. Nevertheless, the views are sufficiently strange to those brought up into the Cartesian way of thinking that at first sight Dreyfus appears to be mystical, or anti-scientific. This is not the case.

Another view of cognition from a phenomological or Heideggerian perspective is given in (Winograd and Flores, 1986); Winograd was instrumental in some of the classical early GOFAI work, before coming around to a very different viewpoint. A different Heideggerian perspective is given in (Wheeler, 1996) within (Boden, 1996). The relevance of Merleau-Ponty is drawn out in (Lemmen, 1998). A different perspective that is similarly opposed to the Cartesian cut is given in (Maturana and Varela, 1987; Varela et al., 1991). A much more general textbook on robotics that is written from a situated and embodied perspective is (Pfeifer and Scheier, 1999).

Heidegger rejects the simplistic objective view, that the objective physical world is the primary reality that we can be certain of. He similarly rejects the opposite idealistic or subjective view, that our thoughts are the primary reality. Instead, the primary reality is our experience of the world, that cannot exist independently of one or the other. Our everyday practical lived experience, as we reach for our coffee or switch on the light, is more fundamental than the detached theoretical reflection that we use as rational scientists. Though Heidegger himself would not put it this way, this makes sense from a Darwinian evolutionary perspective on our own species. From

need not restrict themselves in this fashion (Harvey et al., 1997).

this perspective, our language using and reasoning powers probably arrived in H. sapiens over just the last million or so years in our 4 billion year evolutionary history, and is merely the thin layer of icing on the cake. From both a phylogenetic and an ontogenetic view, we are organisms and animals first, reasoning humans only later.

As humans, of course, detached theoretical reasoning is one of our hallmarks; indeed the Darwinian view presented in the previous paragraph is just one such piece of reasoning. However practical know-how is more fundamental than such detached knowing-that. This is a complete reversal of the typical approach of a Cartesian cognitive scientist or roboticist, who would attempt to reduce the everyday action of a human (or robot) reaching for the coffee mug into a rational problem-to-be-solved: hence the Cartesian sense-model-plan-act cycle.

The archetypal Heideggerian example is that of hammering in a nail. When we do this normally, the arm with the hammer naturally and without thought goes through its motions, driving the hammer home. It is only when something goes wrong, such as the head of the hammer flying off or the nail bending in the wood, that we have to concentrate and start reflecting on the situation, rationalising what the best plan of action will be. Information processing, knowing-that, is secondary and is built on top of our everyday rhythms and practices of practical know-how - know-how which cannot be reduced to a set of rules that we implement. This is true, Wittgenstein suggests, even for our language skills:

> In general we don't *use* language according to strict rules - it hasn't been taught us by means of strict rules either. (Wittgenstein 1960:25)

For the roboticist, this anti-Cartesian alternative philosophy seems at first sight negative and unhelpful. For everyday robot actions this implies that we should do without planning, without the computational model, without internal representations, but nothing has yet been offered to replace such methods. The two lessons that need to be learnt initially is that cognition is

- Situated: a robot or human is always already in some situation, rather than observing from outside
- and Embodied: a robot or human is a perceiving body, rather than a disembodied intelligence that happens to have sensors.

One nice example of a situated embodied robot is the simple walking machine of McGeer which uses 'passive dynamic walking' (McGeer, 1990;

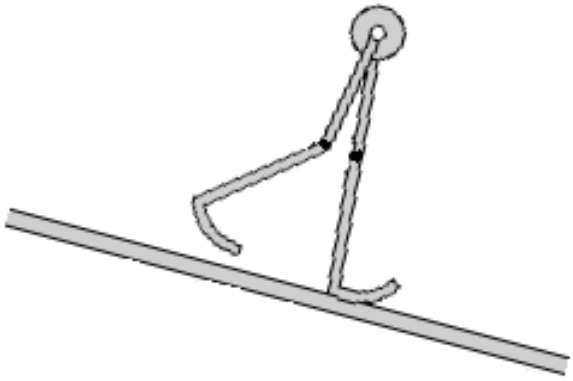

Figure 1: *McGeer's 'passive dynamic walker'.*

McGeer, 1993). This is a two-legged walking robot with each leg made of just an upper and lower limb connected by a knee joint. This knee acts as a human knee, allowing bending freely in one direction but not in the other. At the bottom of each lower limb is a foot shaped in an arc, and the two legs are hinged together at the waist. The dimensions are carefully worked out, but then that is the complete walking robot with no control system needed at all. When it is started off on a gently sloping incline, the legs will walk along in a remarkably natural human-like motion — all the control has come from the natural dynamics, through being situated and embodied.

The walking robot does not take in information through sensors, and does not compute what its current position is and what its next move should be. The designers, of course, carefully computed the appropriate dimensions. In the natural world, organisms and animals have their bodily dimensions designed through natural evolution.

9 The Dynamical Systems alternative

This last example is a robot designed on non-Cartesian principles, using an alternative view that has gained favour in the last decade within AI circles, though its origins date back at least to the early cybernetics movement. One description of this is the Dynamical Systems view of cognition:

> ...animals are endowed with nervous systems whose dynamics are such that, when coupled with the dynamics of their bodies and environments, these animals can engage in the patterns of

> behavior necessary for their survival. (Beer & Gallagher 1992: 91)

At this stage we downgrade the significance of *intelligence* for AI in favour of the concept of *adaptive behaviour*. Intelligence is now just one form of adaptive behaviour amongst many; the ability to reason logically about chess problems may be adaptive in particular refined circles, but the ability to cross the road safely is more widely adaptive. We should note the traditional priorities of AI: the computationalists' emphasis on reasoning led them to assume that everyday behaviour of sensorimotor coordination must be built on top of a reasoning system. Sensors and motors, in their view, are 'merely' tools for information-gathering and plan-execution on behalf of the central executive where the real work is done. Many proponents of an alternative view, including myself, would want to turn this on its head: logical reasoning is built on top of linguistic behaviour, which is built on prior sensorimotor abilities. These prior abilities are the fruit of billions of years of evolution, and language has only been around for the last few tens of thousands of years.

A dynamical system is formally any system with a finite number of state variables that can change over time; the rate of change of any one such variable depends on the current values of any or all of the variables in a regular fashion. These regularities are typically summed up in a set of differential equations. A Watt governor for a steam engine is a paradigmatic dynamical system (van Gelder, 1992), and we can treat the nervous system plus body of a creature (or robot) as one also. The behaviour of a dynamical system such as the governor depends also on the current value of its external inputs (from the steam engine) which enter the relevant differential equations as parameters. In a complementary way, the output of the governor acts as a parameter on the equations which describe the steam engine itself as a dynamical system. One thing that is very rapidly learnt from hands-on experience is that two such independent dynamical systems, when coupled together into (e.g.) steam-engine-plus-governor treated now as a single dynamical system, often behave in a counterintuitive fashion not obviously related to the uncoupled behaviours.

Treating an agent — creature, human or robot — as a dynamical system coupled with its environment through sensors and motors, inputs and outputs, leads to a metaphor of agents being *perturbed* in their dynamics through this coupling, in contrast to the former picture of such agents *computing* appropriate outputs from their inputs. The view of cognition entailed by this attitude fits in with Varela's characterisation of cognition as *embodied action*:

> By using the term *embodied* we mean to highlight two points: first, that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological and cultural context. By using the term *action* we mean to emphasise once again that sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition. Indeed, the two are not merely contingently linked in individuals; they have also evolved together. (Varela et al., 1991: 172–173)

10 Evolutionary Robotics and Behaviourism

Moving from natural agents to artificial robots, the design problem that a robot builder faces is now one of creating the internal dynamics of the robot, and the dynamics of its coupling, its sensorimotor interactions with its environment, such that the robot exhibits the desired behaviour in the right context. Designing such dynamical systems presents problems unfamiliar to those who are used to the computational approach to cognition.

A primary difference is that dynamics involves time, real time. Whereas a computation of an output from an input is the same computation whether it takes a second or a minute, the dynamics of a creature or robot has to be matched in timescale to that of its environment. A second difference is that the traditional design heuristic of *divide and conquer* cannot be applied in the same way. It is not clear how the dynamics of a control system should be carved up into smaller tractable pieces; and the design of any one small component depends on an understanding of how it interacts in real time with the other components, such interaction possibly being mediated via the environment. This is true for behavioural decomposition of control systems (Brooks, 1999) as well as functional decomposition. However, Brooks' subsumption architecture approach offers a different design heuristic: first build simple complete robots with behaviours simple enough to understand, and then incrementally add new behaviours of increasing complexity or variety, one at a time, which subsume the previous ones. Before the designer adds a new control system component in an attempt to generate a new behaviour, the robot is fully tested and debugged for its earlier behaviours; then the new component is added so as to keep to a comprehensible and tractable minimum its effects on earlier parts.

This approach is explicitly described as being inspired by natural evolu-

tion; but despite the design heuristics it seems that there is a practical limit to the complexity that a human designer can handle in this way. Natural Darwinian evolution has no such limits, hence the more recent moves towards the artificial evolution of robot control systems (Harvey et al., 1997).

In this work a genetic encoding is set up such that an artificial genotype, typically a string of 0s and 1s, specifies a control system for a robot. This is visualised and implemented as a dynamical system acting in real time; different genotypes will specify different control systems. A genotype may additionally specify characteristics of the robot 'body' and sensorimotor coupling with its environment. When we have settled on some particular encoding scheme, and we have some means of evaluating robots at the required task, we can apply artificial evolution to a population of genotypes over successive generations.

Typically the initial population consists of a number of randomly generated genotypes, corresponding to randomly designed control systems. These are instantiated in a real robot one at a time, and the robot behaviour that results when placed in a test environment is observed and evaluated. After the whole population has been scored, their scores can be compared; for an initial random population one can expect all the scores to be abysmal, but some (through chance) are less abysmal than others. A second generation can be derived from the first by preferentially selecting the genotypes of those with higher scores, and generating offspring which inherit genetic material from their parents; recombination and mutation is used in producing the offspring population which replaces the parents. The cycle of instantiation, evaluation, selection and reproduction then continues repeatedly, each time from a new population which should have improved over the average performance of its ancestors. Whereas the introduction of new variety through mutation is blind and driven by chance, the operation of selection at each stage gives direction to this evolutionary process.

This evolutionary algorithm comes from the same family as Genetic Algorithms and Genetic Programming, which have been used with success on thousands of problems. The technique applied to robotics has been experimental and limited to date. It has been demonstrated successfully on simple navigation problems, recognition of targets, and the use of minimal vision or sonar sensing in uncertain real world environments (Harvey et al., 1997; Thompson, 1995). One distinguishing feature of this approach using 'blind' evolution is that the resulting control system designs are largely opaque and incomprehensible to the human analyst. With some considerable effort simple control systems can be understood using the tools of dynamical systems theory (Husbands et al., 1995). However, it seems inevitable that, for the same reasons that it is difficult to design *complex* dynamical systems, it is

also difficult to analyse them.

This is reflected in the methodology of Evolutionary Robotics which, once the framework has been established, concerns itself solely with the behaviour of robots: "if it walks like a duck and quacks like a duck, it is a duck". For this reason we have sometimes been accused of being 'the New Behaviourists'; but this emphasis on behaviour assumes that there are significant internal states[3], and in my view is compatible with the attribution of adaptive intelligence. A major conceptual advantage that Evolutionary Robotics has over classical AI approaches to robotics is that there is no longer a mystery about how one can 'get a robot to have needs and wants'. In the classical version the insertion of a value function `robot_avoid_obstacle` often leaves people uncomfortable as to whether it is the robot or the programmer who has the desires. In contrast, generations of evolutionary selection that tends to eliminate robots that crash into the obstacle produces individual robots that do indeed avoid it; and here it seems much more natural that it is indeed the *robot* which has the desire.

11 Relativism

I take a Relativist perspective, which contrary to the naive popular view does not imply solipsism, or subjectivism, or an anything-goes attitude to science. The history of science shows a number of advances, now generally accepted, that stem from a relativist perspective which (surprisingly) is associated with an objective stance toward our role as observers. The Copernican revolution abandoned our privileged position at the centre of the universe, and took the imaginative leap of wondering how the solar system would look viewed from the Sun or another planet. Scientific objectivity requires theories to be general, to hold true independently of our particular idiosyncratic perspective, and the relativism of Copernicus extended the realm of the objective. Darwin placed humans amongst the other living creatures of the universe, to be treated on the same footing. With Special Relativity, Einstein carried the Copernican revolution further, by considering the viewpoints of observers travelling near to the speed of light, and insisting that scientific objectivity required that their perspectives were equally privileged to ours. Quantum physics again brings the observer explicitly into view.

[3]Not 'significant' in the sense of representational — internal states are mentioned here to differentiate evolved dynamical control systems (which typically have plenty of internal state) from those control systems restricted to feedforward input/output mappings (typical of 'reactive robotics').

Cognitive scientists must be careful above all not to confuse objects that are clear to them, that have an objective existence for them, with objects that have a meaningful existence for other agents. A roboticist learns very early on how difficult it is to make a robot recognise something that is crystal clear to us, such as an obstacle or a door. It makes sense for us to describe such an object as 'existing for that robot' if the physical, sensorimotor, coupling of the robot with that object results in robot behaviour that can be correlated with the presence of the object. By starting the previous sentence with "It makes sense for us to describe ..." I am acknowledging our own position here acting as scientists observing a world of cognitive agents such as robots or people; this objective stance means we place ourselves outside this world looking in as godlike creatures from outside. Our theories can be scientifically objective, which means that predictions should not be dependent on incidental factors such as the nationality or location or star-sign of the theorist.

When I see a red sign, this red sign is an object that can be discussed scientifically. This is another way of saying that it exists for me, for you, and for other human observers of any nationality; though it does not exist for a bacterium or a mole. We construct these objects from our experience and through our acculturation as humans through education[4]. Just as our capacity for language is phylogenetically built upon our sensorimotor capacities, so our objects, our scientific concepts, are built out of our experience. But our phenomenal experience itself cannot be an objective thing that can be discussed or compared with other things. It is primary, in the sense that it is only through having phenomenal experience that we can create things, objective things that are secondary.

12 Conclusion

Like it or not, any approach to the design of autonomous robots is underpinned by some philosophical position in the designer. There is no philosophy-free approach to robot design — though sometimes the philosophy arises through accepting unthinkingly and without reflection the approach within which one has been brought up. GOFAI has been predicated on some version of the Cartesian cut, and the computational approach has had enormous success in building superb tools for humans to use — but

[4]It makes no sense to discuss (...for *us humans* to discuss ...) the existence of objects in the absence of humans. And (in an attempt to forestall the predictable objections) this view does *not* imply that we can just posit the existence of any arbitrary thing as our whim takes us.

it is simply inappropriate for building autonomous robots.

There is a different philosophical tradition which seeks to understand cognition in terms of the priority of lived phenomenal experience, the priority of everyday practical know-how over reflective rational knowing-that. This leads to very different engineering decisions in the design of robots, to building *situated* and *embodied* creatures whose dynamics are such that their coupling with their world leads to sensible behaviours. The design principles needed are very different; Brooks' subsumption architecture is one approach, Evolutionary Robotics is another. Philosophy does make a practical difference.

Acknowledgments

I thank the EPSRC and the University of Sussex for funding, and Shirley Kitts for philosophical orientation.

References

Abeles, M. (1982). *Local Cortical Circuits, an Electrophysiological Study.* Springer-Verlag.

Appel, K. and Haken, W. (1989). Every planar map is four colorable. *American Mathematical Society, Contemporary Mathematics*, 98.

Beer, R. D. and Gallagher, J. C. (1992). Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1(1):91–122.

Boden, M. A. (1990). *The Philosophy of Artificial Intelligence* Oxford University Press.

Boden, M. A. (1996). *The Philosophy of Artificial Life* Oxford University Press.

Brooks, R. (1999). Cambrian Intelligence: the Early History of the New AI. MIT Press, Cambridge MA.

Dreyfus, H. L. (1965). Alchemy and Artificial Intelligence. RAND Corporation Paper P-3244, December 1965.

Dreyfus, H. L. (1972). What Computers Can't Do: a Critique of Artificial Reason. Harper and Row, New York.

Dreyfus, H. L. and Dreyfus, S. E. (1986). Mind Over Machine.

Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. (1997). Evolutionary robotics: the Sussex approach. *Journal of Robotics and Autonomous Systems* v. 20 (1997) pp. 205-224.

Husbands, P., Harvey, I., and Cliff, D. (1995). Circle in the round: State space attractors for evolved sighted robots. *Journal of Robotics and Autonomous Systems. Special Issue on "The Biology and Technology of Intelligent Autonomous Agents"*, 15:83–106.

Lemmen, R. (1998). Towards a Non-Cartesian Cognitive Science in the light of the philosophy of Merleau-Ponty. DPhil Thesis, University of Sussex.

Malsburg, C. von der, and Bienenstock, E. (1986). Statistical coding and short-term plasticity: a scheme for knowledge representation in the brain. In Bienenstock, E., Fougelman-Soulie, F., and Wiesbuch, G. (eds.) *Disordered Systems and Biological Organization*. Springer-Verlag.

Marr, D. C. (1977). Artificial Intelligence: a personal view. *Artificial Intelligence*, 9:37-48.

Maturana, H. and Varela, F. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Shambhala Press, Boston.

McGeer, T. (1990). Passive dynamic walking. *Int. J. Robotics Research*, 1990(9)2:62-82.

McGeer, T (1993). Dynamics and Control of Bipedal Locomotion. *J. Theor. Biol.*, 1993(163), 277-314.

Pfeifer, R., and Scheier, C. (1999). *Understanding Intelligence*. MIT Press.

Thompson, A. (1995). Evolving electronic robot controllers that exploit hardware resources. In Morán, F., Moreno, A., Merelo, J.J. and Chacón, P. (eds.) *Proceedings of Third European Conference on Artificial Life*. Springer-Verlag, pp. 640–656.

van Gelder, T. (1992). What might cognition be if not computation. Technical Report 75, Indiana University Cognitive Sciences. Reprinted in *Journal of Philosophy* 92:345–381 (1995).

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind*. MIT Press.

Wheeler, M. (1996). From Robots to Rothko: The Bringing Forth of Worlds. In Boden, M. (ed.) *The Philosophy of Artificial Life*. Oxford University Press.

Winograd, T. and Flores, F. (1996). *Understanding Computers and Cognition: A New Foundation for Design*. Ablex Publishing Corporation, New Jersey.

Wittgenstein, L. (1960) *The Blue and Brown Books* Basil Blackwell, Oxford.