Evolution and the Origins of the Rational

Inman Harvey
Evolutionary and Adaptive Systems Group
COGS/Dept. of Informatics
University of Sussex
Brighton BN1 9QH, UK
inmanh@cogs.susx.ac.uk

Unshared assumptions

A civilized disagreement should ideally start with a set of recognized, shared assumptions. This at least forms the basis for comprehension of the context for the argument, of the terms used and their baggage of implicit premises. Discussion between academics rarely meets this ideal, and even more rarely when they come from different disciplines.

I and my colleagues are in the business of investigating fundamental ideas about cognition by creating working artifacts that demonstrate basic cognitive abilities. So to that extent we count as cognitive scientists, and indeed our work can be seen as being at one end of the Artificial Intelligence (AI) spectrum; though, for reasons that may become apparent, we are wary of this label and the baggage it carries. We work with neuroscientists, ethologists, evolutionary theorists and roboticists, so the struggle to communicate across disciplinary boundaries is all too familiar to us. Our projects require us to take a philosophical stance on what cognition is all about, but our desired goal is not so much a theory as a working product, a working robot (or a principled computer simulation of a robot) that demonstrates some cognitive ability: intentionality, foresight, ability to learn, to communicate and cooperate, or rational behaviour.

So the goal is to create physical mechanisms, if not of flesh and blood then made of plastic and transistors, that demonstrate behavior describable in the language of the mind. Of course, if some specific mechanism can generate behaviour X, this does not necessarily imply that humans or animals use the same or even a similar mechanism. Nevertheless, such a demonstration is an existence proof: "See, we can demonstrate phenomenon X with no more than these physical components, with nothing up our sleeves. So any claim that some other ingredient is **essential** for X has been disproved."

This 'philosophy of mind with a screwdriver' is, we claim, more challenging than the armchair version. A working robot becomes a puppet displaying the philosophical stance of its creators, and is an invaluable tool for making explicit the assumptions used, and for exposing the assumptions of those who doubt its abilities. In fact we try and minimize the assumptions and preconceptions used when designing our robots or simulated agents, by using an Evolutionary Robotics (ER) approach: we use an artificial analogue of natural Darwinian evolution to design the robot/agent mechanisms.

One assumption that we most emphatically do not share -- with many philosophers of mind, cognitive scientists and AI practitioners – is that physical mechanisms and mental mechanisms are pretty much the same, that the architecture of the brain is isomorphic or even similar to the architecture of the mind. The mechanism of a clock can be specified in terms of a pendulum, weights, cogs, levers and springs, and carries with it no mention of timekeeping, even though that is the role of the clock. Likewise, we are interested in specifying at the physical level what constituents are sufficient for assembling into an artificial agent that displays cognition, and it is pointless to try to label any of these physical parts with mental or cognitive terms. This pernicious habit is regrettably rife in AI, but one of the core insights of a Dynamical Systems (DS) approach to cognition, as discussed below, is that a state of mind should not be equated with a physical state of the brain; no more than a 'state of running fast' in a clock can be equated with a state of one or many, or even all, the parts of its clockwork.

This confusion largely arises because the term 'state' has different usages within different disciplines. For someone using a DS approach, the instantaneous state of a physical mechanism is specified by a vector: the current physical values of all the relevant physical variables of the system, including the instantaneous angle and angular velocity of the pendulum and of all the cogs in a clock, the electrical potentials and chemical concentrations at all relevant parts of a real brain, the sensor and motor configurations and all the internal instantaneous activations of a robot brain. In this sense, the state of a physical mechanism is a vector of real numbers that is continuously varying (unless the clock, the brain or the robot is effectively dead). By contrast, a state of mind such as belief, knowledge, fear, pain is something that is extended in time.

When challenged, an identity theorist who equates mental states with brain states turns out not to be using this vector terminology to describe a state of the brain. The claim usually takes the form of some subpart, perhaps some module of the brain, being capable of maintaining over the relevant extended period of time a stable value, or stable pattern of values, which is causally correlated to the mental state – independently of whatever values the other variables in the brain might take. In my preferred terminology this is not a 'brain state', but rather a 'subset of possible states of a subpart of the brain'. Even with this revised terminology, we can demonstrate robots that display mental states that have no such identifiable physical correlate within the (artificial) brain; only when we take into account the physical state of the environment, and the agent's current physical interaction with the environment, can the full explanation be given of how the cognitive phenomena arise from the physical mechanisms.

So this preamble is by way of a warning that the assumptions used here may be rather different from your own; and that we may reject some of the standard arguments against the philosophical positions we take, not because we disagree with the form of the argument but because we reject some of the assumptions on which they are based. Firstly, we should discuss what we mean by Cognition and the Rational.

Chauvinism

The English are notoriously bad at learning foreign languages. In fact, we don't see the point of foreign languages at all -- why should the French call a cow by some arbitrary word such as 'vache' when clearly it is a **cow**? God, after all, is an Englishman, and clearly the natural language of the universe is English.

Philosophers tend to be similarly chauvinistic -- but species-chauvinist rather than nation-chauvinist. By 'rationality' we usually mean human rationality, indeed usually 20th or 21st century western educated human rationality; and too often we assume that this rationality is some God-given eternal yardstick. Just as the English question the wisdom of, indeed the need for, any other language, so the philosopher takes our human-specific rationality as the only perspective that makes sense.

There is some virtue in this; after all if we speak we must speak in the language we have available, and if we are trying to make sense of the world then we are trying to make sense from our own human perspective. But I suggest that many of the problems and confusions in cognitive science come from this unthinking chauvinism, and above all when we consider the origins of the rational.

There are at least two meanings for the term rational. On the one hand, we apply the word literally to human acts of reasoning, of logical or systematic reasoning; we can understand the steps in the proof of a theorem, and we can understand how the smooth curves of an aeroplane have arisen from thoughtful consideration by aeronautical engineers. On the other hand, we apply the word metaphorically to articles where we can, a posteriori, reason that form follows function as if it had been rationally designed; the smooth curves of a dolphin can be rationalised as a near-optimal design for swimming -- they are a rational shape. A major revelation of Darwinian evolution was that 'rational' design can be the product of the mechanical unthinking process of heredity, variation and natural selection.

With this Darwinian revelation in mind, we should always be careful, when talking about any living creatures including humans, to distinguish between 'rational' in the sense of explicit reasoning, and 'rational' in the metaphorical or 'as-if' sense that the theory of evolution now gives us. Let us summarise this briefly.

A Brief History of Life

Here is a simplified version of the typical working assumptions of many biologists today. Many details are glossed over, there is room for disagreement on many of the specifics, but the overall picture will be generally shared.

Around 4 billion years ago on this planet, some relatively simple living creatures arrived or arose somehow. There are several current theories as to how this might have happened, but the details do not matter for our present purposes. What does matter is that these creatures reproduced to make offspring that were similar but not necessarily identical to their parents; and the variations in the offspring affected their likelihood of surviving and producing fresh offspring in their turn. The likelihood of surviving depended on their ability to extract nutrients from their surroundings and avoid harm. Those that happened to do better than others, in their particular circumstances, would tend to be the parents of the next generation; so that the children tended to inherit their characteristics from relatively successful parents.

Fitness depended on the particular requirements for nutrients, and the particular type of environmental niche. Over many generations, this Darwinian evolution working on the natural blind variation resulted in organisms that looked as though they were crafted, were rationally designed for their life style.

The Blind Watchmaker produced 'as-if' rational design, without explicit reasoning. It also produced organisms with wants and needs, and with purposive behaviours that tended to satisfy those needs. We should be suspicious of arguments that produce an "ought" from an "is"; but we can argue that Darwinian evolution shows us just how a "want" can be derived from an "is".

Initially, all such organisms were single cells; the distinction between self and other was clearly defined by the cell membrane. Cells competed against other cells of their own kind
(close relatives) and of other species (distant relatives) for limited resources. At some stage it occasionally became viable for limited coalitions to form. Multicellular organisms 'put all their eggs in one basket' as far as their joint and several fitnesses were concerned. New issues arose of policing the new boundaries between self and other, and ensuring that the commonwealth of interests within the boundary was not disrupted by the cancer of cheating. Life got more complex.

Organism's needs were served by not merely reacting to the immediate, but by anticipating future occurrences. Even plants have nervous systems and can learn, For instance sun-seeking flowers such as Malvastrum Rotundifolium, in order to maximise the time they spend soaking up the warmth face-on to the sun, return overnight to face the expected direction of the dawn. This direction changes with the seasons, and experiments show that it takes just a couple of days for some plants to learn where to expect a changed sunrise. Plant cognition for a plant's lifestyle.

Some organisms are more mobile than others in seeking out sustenance; animals prey on plants and other animals. More sophisticated nervous systems are required to deal with the faster and more varied world that this change of lifestyle implies. For some animals, recognition of places, and of other individuals within ones species, becomes important. Social behaviour requires knowing who to trust and who to distrust. The sophisticated manipulation of other creatures displayed by bacteria and plants becomes even more sophisticated and faster.

After 4 billion years of evolution, for a brief period of a few million years one particular slightly different species flourished before being ousted by other species with different interests. Humans shared most of their cognitive faculties with their bacterial, plant and animal relatives, but adjusted to their particular lifestyle. They had hunger and fear, they searched for food and avoided danger. They could recognise what was good and bad for them, they could learn from experience and alter their behaviour accordingly. They had a rich social life and good memory for faces.

I am emphasising here the seamless continuity of human cognition with the cognition of all biological organisms. The differences should not blind us to the fact that most of our cognitive faculties existed in other creatures prior to the arrival of humans on the scene. But then, within this context, we should acknowledge the peculiar differences that distinguish humans from other species.

What Makes Humans Different

It just so happened that the strategic direction evolution chanced on for this species was towards increased social interaction and manipulation. Division of labour had been seen in many insect societies, but was carried much further amongst humans. Though division of labour has potential great benefits for the participating individuals, it requires trust and mechanisms for retaliating against or shunning those who betray that trust.

Consensual manipulation of others became important, and the relatively simple methods of communication used by plants and animals became extended into a complex system of stereotypical sounds. As well as simple commands and warnings, intentions and coordination of future actions could be conveyed. Different individuals, with different roles in some coordinated action, could convey information and negotiate as to possible strategies.

Discussion, and then the ability within discussion to anticipate events and other peoples' situations, became a defining human characteristic. First came the birth of language, and then the birth of thinking; for the language of thought is language.

One subset of language is the language of rational thought. With this, we can reason about possible situations that are not present; by developing mathematics and logic we can reason about whole classes of events.

Completing the Circle of Rationality

So we should recognise that we are one biological species amongst many. Evolution has crafted the design, the behaviour, the cognitive faculties of all species to be generally appropriate to the interests of those organisms given their particular lifestyles. At some stage the peculiar human faculty of rational thought, made possible by language, arrived on the scene.

We can rationally, consciously and with explicit intent design artifacts to do a job. It was Darwin's brilliant insight that the irrational, unconscious and unintended forces of natural evolution can design adapted organisms that look as if they are rationally designed; alternatively, one can say that their designs are rational -- in what is now perhaps a metaphorical sense of the word.

So I want to distinguish between two different senses of the word 'rational': a primary sense, shall we say rational_a, referring to the explicit reasoning that we humans with our languages are exclusively capable of: and a secondary derivative sense rational_b, where we appreciate Darwin's insight that naturally evolved designs can be near-optimal for some function without a rational_a designer.

Confusingly, in terms of evolution rational_b designs occurred historically prior to the emergence in humans of rationality_a; and indeed I have spelt out above one possible rationalisation_a as to just how rationality_a arose because it was rational_b for the lifestyle that humans came to take up!

Studies of Minimal Cognition

Many of my colleagues working in cognitive science and AI are species-chauvinistic. They are solely concerned with that part of human cognition that is exclusively human, and fail to see the relevance of the related cognitive abilities of our animal, plant and bacterial cousins and ancestors.

Within our Evolutionary and Adaptive Systems Group, however, we do explicitly follow up these relationships. On the basis that one should crawl before one runs, we are interested in exploring minimal levels of cognition in adaptive organisms, where 'adaptive' refers to behaviour that is rational_b in the face of changing circumstances.

We use the tools of Artificial Life to synthesise artificial creatures -- simple actual autonomous robots or simulated versions of them -- to test ideas on how rational_b behaviour is generated. By autonomous we mean that, as far as is feasible, the artificial nervous system should be self-contained and these creatures should 'do their own thing' without the experimenter making internal adjustments.

Philosophy of Mind

We use synthesised, model creatures sometimes to investigate how some very specific feats of animal cognition are done: the visual navigation of ants and bees, for instance. But also some of our experiments are for exploration of more general issues of cognition, with wider applicability across many or perhaps even all organisms, both actual and potential.

When we work with synthesised creatures we know exactly what is and is not there; there can be no vitalistic forces hidden up our sleeves. If we can demonstrate intentional behaviour, foresight, learning, memory, desire, fear, re-identification of temporarily missing objects, social coordination, language use ....; if we can demonstrate any of these in a simulated creature whose inner workings are visible to us, then we have an existence proof for one possible way that this might have been achieved in the natural world. It may not be the same way, but at a minimum this can perhaps help dispel some of the mysteries that have puzzled people for ages.

Some of these mysteries fall into the remit of Philosophy of Mind, and the experiments that we do have some resemblance to thought experiments. But there is this important difference: our experiments do not survive unjustified or careless underlying assumptions, they can fail. This is the crucial test for philosophy of mind with a screwdriver.

Wittgenstein famously saw the aim of philosophy as ‘shewing the fly the way out of the fly-bottle’. Confused buzzing within the philosophy of mind includes such questions as: just how can the physical substance of brain and body generate mental phenomena such as intentional behaviour? Is some vital extra ingredient needed? When a human or other creature takes account of some object in its present environment, or uses foresight to take account of an object that is currently out of sight, how do the physical mechanisms of the nervous system relate to such an object?

How can social coordination arise, how can we understand the origins of language in terms of a physical system made of meat and nerve fibres, or metal and silicon?

The flybottle that has trapped so many people in these areas is the genuine puzzlement many have about the relationship between the physical and the mental. The perspective I take is that there are (at least) two different levels of description appropriate for humans, animals and other agents, including robots. We can describe the physical substrate of the agent: this is the language of the doctor, the anatomical neuroscientist, the robot engineer. Or we can describe an agent in behavioural and intentional terms, in 'mentalese' such as goals and desires, terms that are only appropriate for living creatures that have primary goals of survival and secondary derived goals; these mentalese terms may (cautiously) be extended where appropriate to artificial creatures.

The fly-bottle focused on here is: how can we reconcile these two languages of description? A common cause of confusion arises from careless use of language, from smuggling in mentalese terms such as 'representation' into what should be exclusively physical descriptions; the Physical Symbol Systems approach of classical AI being a common source of this confusion.

In Beatrix Potter's children's story, it was thought to be sufficient explanation for why lettuce made rabbits sleepy to announce that they contained something soporific. Hopefully philosophers are less easily satisfied; it should be obvious that one cannot attempt to explain the relationship between the mental and physical by smuggling mental terms into physical descriptions. So a first recommendation for the way out of this fly-bottle is to rigorously police the use of language: mental descriptions can and should be in mentalese, but physical-level descriptions should use no such terms.

There is a reason why we should be tempted into the trap of using mentalese in our physical descriptions. If we believe some version of the tale above on the origins of language, this was partially, perhaps primarily, spurred on by the need to understand, plan and manipulate human relationships. An explanation typically works by restating the complex and unfamiliar in terms of the familiar and unquestioned. So we will explain central heating systems in terms of homunculi: “the thermostat is like a little person assessing the temperature, and passing on a message down the wires to the box in the control system that acts like a little human manager: in turn, this sends the equivalent of a human command to the switch that turns the boiler on”. This is perfectly aceptable, human-friendly language of explanation that we use everyday; it is normally sensible and the metaphors have much to recommend them. Just in this one special case, of elucidating the relationship between mentalese language of behaviour and the physical language of mechanism, it is a crucial trap for the careless fly (see how naturally we use a fly-homuncular metaphor to get the point across).

Cartesian or Classical approaches to Robotics

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. 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.

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.

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 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, one should 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?

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.

The Dynamical Systems approach to Cognition

In the following section we shall be outlining the strategy of Evolutionary Robotics as a design methodology for putting together robot nervous systems from a shelf-full of available components. But first we must establish just what sorts of components are necessary and sufficient – and the components will not be intended to carry any representations.

The nervous system of an animal is an organized system of physical components such as neurons, their connecting axons and dendrites, and their substrate, with electrical and chemical activity swirling around. The picture that neuroscientists give us is still changing; it is only in the last decade or so that the significance of chemical transmission as well as electrical transmission between neurons has been noted as significant. But the universal working assumption is that in principle there is a finite (though extremely large) number of physical variables that could be picked out as relevant to the workings of the machinery of the brain; and these variables continuously interact with each other according to the laws of physics and chemistry.

When we have a finite number of variables and can in principle write down an equation for each one, stating how its rate of change at any instant can be given by a formula related to the instantaneous values of itself and some or all of the other variables, then formally speaking we have a Dynamical System. In the case of a nervous system, the variables are not only internal ones, but also include sensory inputs and motor outputs, the interactions with the world around it.

So for our robots we pick appropriate sensors and motors for the task, bearing in mind that it is unwise to treat, for instance, an infra-red transceiver as a distance-measurer; rather, it is a transducer that outputs a voltage that depends on a variety of factors including the distance from the nearest object, the ambient light levels, the material that a reflective surface is made from. An infra-red detector may well indeed be useful for a robot that needs to avoid obstacles, but it would be a mistake to label the output as “distance-to-wall”. This would be a term in ‘mentalese’ rather than a neutral physical term.

As for the internal components, for some specific ER experiments these have been components of standard electronic circuits as packaged in a Field Programmable Gate Array. But for reasons of practicality, in many experiments we use Artificial Neural Networks (ANNs) as simulated in real time with the aid of a computer. In particular one favourite class of ANNs is the CTRNN, Continuous Time Recurrent Neural Network (Beer 1995). This is a potentially fully-connected network of real time leaky integrators with specified temporal constants, and is an archetypal dynamical system. The class of CTRNNs has the useful property of universal approximation to any smooth dynamical system (Funahashi and Nakamura 1993); in other words, given any DS where the variables change smoothly, we can in principle find a CTRNN, with enough nodes, that will approximate its dynamical behaviour to any desired degree of accuracy.

Evolutionary Robotics

The problem of AI is just how to build systems that generate adaptive or intelligent behaviour. With an evolutionary perspective, it makes sense to start small and simple first, so we look at minimally cognitive artificial agents; in particular with rational_b adaptive behaviour as a more fundamental starting place than the rational_a behaviour of human intelligence.

How can we design the control mechanisms which can produce adaptive behaviour in synthetic creatures? We must bear in mind that the mechanism must be described in non-intentional language if this is going to give us any insight into the relationship between the mental and the physical.

Braitenberg's Vehicles are a very simple example of what we are seeking. We can describe a very simple circuit where Left and Right photocells, mounted on the front of a simple toy vehicle, are connected respectively to the Right and Left motors driving the wheels on each side. In the simplest version, the circuit drives each wheel slowly forward in the absence of any light; any increase in the light-level reaching either photoreceptor results in increased speed on the attached wheel. I have described this 'nervous system' in non-mentalese, yet we can see quite easily that this results in light-seeking behaviour in the vehicle. Further, experiment shows that this behaviour is adaptive, in the sense that the photovore behaviour will adapt the motion to continually chase a moving light-target. The observer watching this behaviour just naturally describes it in intentional language. This simple example helps to resolve worries about how such relatively simple organisms as bacteria can display similarly goal-directed behaviour; there is no need to call on explicit conscious intentions.

How do we scale up from this to more sophisticated examples? One approach is to follow the course of evolution fairly literally; this gives us the relatively new field of Evolutionary Robotics, as a methodology for designing artificial creatures using artificial Darwinian evolution.

Suppose that we wish to produce a creature that will repeatedly seek a light target, but also learn what landmarks will aid it in the search. Then as Evolutionary Roboticists we set up a test environment, where robots can be evaluated and scored on their fitness at this task. We then work with a population of robots that have various designs of nervous system architecture; or more practically, we usually work with one robot and a population of possible architectures. The components of the nervous system are real or idealized physical components equivalent to the pendulums, cogs and levers of a clockmaker; we shall return in more detail to this below. In our role as the Creators of this artificial world, we specify some appropriate mapping between genotypes, strings of artificial DNA that may well be simply composed of 0s and 1s, and phenotypes by which we mean the actual way in which the nervous system is assembled from the available components.

At the simplest level, the genotype may be (when translated) in effect a blueprint for assembling the nervous system. With more sophisticated mappings, it may act more like a recipe ‘for baking a cake’, directing and influencing (perhaps in league with environmental influences) the final architecture without actually explicitly specifying its form. Whichever form of mapping we choose to use, the result should be that the

search through the space of possible architectures is paralleled by an equivalent search through the space of possible genotypes, of artificial DNA. Indeed, since these genotypes allow inheritance of genetic material from robot parents selected for their fitness at a task, and mutations to the genotype – a few random changes to the 0s and 1s – allow for variation, we have all the necessary ingredients for Darwinian evolution: Heredity, Variation and Selection.

The Evolutionary Procedure

Artificial evolution typically consists of repeated rounds, or generations, of testing a population of candidate designs, and selecting preferentially the fitter ones to be parents of the next generation. The next generation consists of offspring that inherit the genetic material from their selected parents; much the same as farmers have been doing for thousands of years in improving their crops and their livestock. The initial population is based on completely random genotypes of artificial DNA, so the only direction given to the evolutionary design process is the indirect pressure of selection. The genotypes in following generations are usually mixed through sexual recombination, and further varied through mutations, so as to introduce further variety for selection to choose from.

As far as possible this is an automated, hands-off process. The human designer's input is limited to the mapping from genotype to phenotype, and the selection process that allocates fitness scores to each robot. This depends on the cognitive ability that is desired, and usually requires careful thought. To give a simple example, if one wishes to craft a fitness function intended to promote the movement of a robot across a crowded floor without bumping into anything, then one might be tempted to judge the fitness by how far the robot travels in a fixed time. Although high scores will be achieved through evolution, the result may well be disappointing as typically such scores will be gained by rapid rotation around a tight circle; it turns out to be necessary to craft the fitness function so as to give more credit for (fairly) straight movement and less for tight turns.

Although the human designer has set up the scenario intended to select, over many generations, for the desired behaviour, it is important to note two things. Firstly, no individual robot is given any feedback as to how well its behaviour is accumulating fitness – so later generations only succeed if it is in their inherited 'genetic nature' to behave appropriately. Secondly, more often than not the finally evolved robot nervous system is complex and opaque, it is difficult and maybe impossible to analyse just how the job is done.

Some experiments

The techniques of ER have been developed since the beginning of the 1990s, and by now there are probably thousands of evolved robot designs, both in simulation and with real robots. Three major centres are Case Western Reserve University, where Beer and colleagues perform simulation experiments in 'minimal cognition' (Beer 2000); EPFL in Lausanne, Switzerland, where Floreano and colleagues eveolve control systems for real robots (Floreano and Urzelai 2000); and our own group at Sussex, where we work both with simulations and real robots (Harvey et al. 1997).

The types of cognitive behaviour that we can demonstrate in robots through these means start at the very simple and then work slowly up. We can generate visual awareness and recognition of particular distant objects or goals, and the ability to navigate towards them while avoiding obstacles. There have been studies on the origins of learning, where an agent needs to learn and relearn about correlations (between a visual landmark and its goal) that change at unexpected intervals (Tuci et al. 2003). There has been a successful attempt to recreate in a robot the equivalent to our human ability to adjust to wearing inverting glasses – after an extended period of time, weeks in the case of humans, the world that was seen as upside-down becomes familiar enough to allow us to get around and navigate successfully (Di Paolo 2000). Teams of identical robots have been evolved so as to effectively negotiate between each other their different roles in a joint operation (Quinn et al. 2003).

All these behaviours display the hallmarks of intentionality at different levels of complexity, despite the fact that the robot nervous system or control system is nothing more than a dynamical system cobbled together through evolution. The selection pressures were designed through consideration of the intentional behaviour required, in 'mentalese' terms; but the genetic specification of the brain is in pure physical terms, the neutral language of DS. Analysis of the evolved networks typically shows no 'perception module' or 'planning module' in the brain, no obvious representational place-holders for internal representations of external objects or events. These experiments give us an existence proof for one possible way of generating these behaviours, and hence act as a challenge to those who claim that brains must be organized in some radically different way in order to produce these effects.

All these experiments demonstrate rational_b adaptive behaviour, but not yet the rational_a behaviour of human intelligence. What we have done is intended as echoing, in simplified form, the evolution of cognition in the relatively early days of life on this planet. Although there is plenty of work to be done before aiming explicitly at human rationality, the experiments in communication and coordination of behaviour between robots is one very small stepping stone aimed in that direction.

To Summarise

I have tried to give a feel for the working philosophy of someone working in new, non-Classical AI, Artificial Life and Evolutionary Robotics. This means rejecting a lot of the philosophical baggage that has been associated with much of AI and Cognitive Science in previous decades. It means continually needing to explain oneself to others who start from a very different set of assumptions.

There is a massive rift in AI and Cognitive Science at the end of the twentieth and beginning of the twenty-first century. I would suggest that this may be a sign of a Copernican revolution finally reaching Cognitive Science, up to a century or more after equivalent revolutions in biology and physics.

The Copernican Revolution

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.

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. 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.

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.

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. Computationalist AI 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 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.

Bibliography

Beer, R.D. (1995). On the dynamics of small continuous-time recurrent neural networks Adaptive Behavior 3, 469-509.

Beer, R.D. (2000). Dynamical approaches to cognitive science. Trends in Cognitive Sciences 4(3):91-99.

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

Di Paolo, E. A., (2000). Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. Proc. of SAB'2000, MIT Press.

Floreano, D. and Urzelai, J. (2000) Evolutionary Robots with on-line self-organization and behavioral fitness. Neural Networks, 13, 431-443.

Funahashi, K. and Nakamura, Y. (1993). Approximation of Dynamical Systems by continuous time recurrent neural networks, Neural Networks 6, 801--806.

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

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

Quinn, M., Smith, L., Mayley, G. and Husbands, P. (2003). Evolving controllers for a homogeneous system of physical robots: Structured cooperation with minimal sensors. Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences, 361:2321-2344.

Tuci, E., Quinn, M., and Harvey, I.: An evolutionary ecological approach to the study of learning behaviour using a robot based model Adaptive Behavior 10(3/4):201-222, 2003.