PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Behaviour: Perception, Action and Intelligence - The View from Situated Robotics [and Discussion]

J. C. T. Hallam, C. A. Malcolm, M. Brady, R. Hudson and D. Partridge

Phil. Trans. R. Soc. Lond. A 1994 **349**, 29-42
doi: 10.1098/rsta.1994.0111

Behaviour: perception, action and intelligence – the view from situated robotics

By J. C. T. Hallam and C. A. Malcolm

Department of Artificial Intelligence, University of Edinburgh, 5 Forrest Hill, Edinburgh EH1 2QL, U.K.

We relate the problems afflicting implementations of classical knowledge-based symbolic systems to theoretical criticisms of the paradigm, and explain why many of those pursuing research programmes designed to avoid these problems and underpinned by models of mind variously described as 'behaviour-based', 'reactive', 'enactive', 'situated', 'embedded' are using robot rather than computer systems as their experimental domain. We argue that mentalistic terms are only applicable to contingent historical agents embedded in the local world with which they interact, and therefore (for example) attempting to implement intelligence, semantics, etc., in a computer system is a doomed enterprise.

1. Introduction

Artificial Intelligence as a science is concerned with understanding how intelligent behaviour is produced. It differs from other branches of cognitive science in its methods. This understanding is developed and proved by making artificial devices which display intelligent behaviour. If we accept, as, for example, Newell & Simon (1976) suggest in their physical symbol system hypothesis, that intelligent behaviour is knowledgeable behaviour, the essential problems being how the knowledge is represented, accessed and reasoned with, then we have the classical paradigm in AI, what Haugeland (1985) has called 'good old fashioned AI' (GOFAI) and Harnad (1989), 'symbolic functionalism'. Here the computer is the experimental tool and the research is done by finding ways of implementing the required knowledge, with its access and inference machinery, in a computer.

Many do not accept that the essential problems behind the production of intelligent behaviour are the representation of knowledge. Some consider that knowledge is only part of the story (see, for example, Dreyfus 1979), and some consider that knowledge and its mechanisms are a formal *description* of the constraints on intelligent behaviour, but have nothing to do with the *production* of the behaviour. It is, of course, often *possible* to produce a certain kind of behaviour by using a formal description of the behaviour as part of the implementation; but often this is a computationally very expensive way of doing this.

If we are AI researchers who do not accept the premises of GOFAI or symbolic functionalism, i.e. the central role of explicit knowledge, then we can no longer rely on purely computer-hosted knowledge-based systems. The most general kind of intelligent device, which makes the least presumptions about the nature of

The catchline throughout this paper should read:
Phil. Trans. R. Soc. Lond. A

mind and intelligence, is the robot, an artificial agent going about its business in the real world.

This is why robotics has such an important role in Artificial Intelligence, and why among today's roboticists are to be found many of those who disagree with the premises of GOFAI. While other subfields of AI attempt to model, and thereby study, particular aspects of cognition, they always permit the resulting model to be sheltered from reality to some extent by its partly indirect interface to the world through the user or investigator. An autonomous robotic system, however, faces the world on its own terms. It attempts without presumptions the complete unadulterated problem: the intelligent behaviour of an autonomous agent.

A recent growth of interest in aspects of intelligent behaviour which may lie outside the knowledge-based paradigm has led to an emphasis on mentality embodied in robots, in which the *embeddedness* of the system in the world is seen as crucial both to the function of the system and to a true scientific understanding of the system. In this paper we attempt to characterize the various strands of work in progress calling themselves variously 'situated', 'behaviour-based', 'enactive', or 'reactive', and to assess their implications for the scientific and philosophical explanation of autonomous intelligence.

2. The classical paradigm

Paradigms are crucial to science. They shape the way we think about the world, determine the questions we consider in research, influence our interpretation of the results and affect the success or otherwise of our efforts. Furthermore, the philosophical and technological assumptions on which they are based are not normally accessible to analysis within the paradigm and are often unconscious ones, as far as many researchers working under the banner are concerned.

Artificial Intelligence is no exception to this rule. Until recently, the major AI paradigm could be characterized by its principal assumption that a (symbolic) description of cognitive processes in a manner independent of the particular hosting hardware was possible, and that such a description would provide the key to the implementation of cognition. Epitomized by the seminal work of Newell & Simon (1976), this approach views cognition as a computational process in which the behaviour of agents is engendered by the rational application of knowledge about the world and the agent.

In this view of things, an agent is engaged in a continual cognitive cycle. Variously called the 'move, stop and think' (in the context of navigation (Hallam 1983)) or the 'sense–model–plan–act' framework (within the context of planning (Brooks 1991)), the cycle consists of the collection of sensory data and its integration into an internal representational model of the world, followed by inspection of that model, determination of what happened and rational selection of the next action(s) of the agent.

One way of thinking about this is in terms of the diagram devised by Rosen (1987), shown in figure 1. The figure depicts graphically the relation between a formal theory and the reality it is intended to model. On the left, we see some process in the real world (this might be as simple as a falling body, or as complex as a human being) whose dynamics are the result of physical causality; on the right, we have a formal model of the process, often constructed in a mathematical language, whose 'dynamics' are the formal rules of inference appropriate to the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

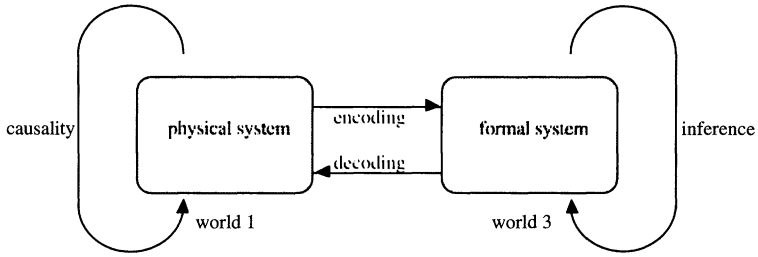

Figure 1. The relation between a real system and a formal model of it.

modelling language used; and between the two we have functions that encode and decode the model, i.e. they map objects and properties in the real world to objects and properties in the formal model, and vice versa. Encoding and decoding are constructed so that the diagram commutes in the mathematical sense; in other words, a decoded deduction from the encoded representation of a physical situation is the causal consequence of that situation. The formal model on the right is the one a scientist might construct to explain the process on the left and to predict its properties or evolution.

To a roboticist considering how to design an intelligent robot, this is a very appealing model of how to think about the world, well known to scientists and well validated by the success of science.

Two further remarks are appropriate before we proceed. First, we should note that the right-hand side of the diagram of figure 1 comprises entities that exist in Popper's world 3, i.e. the communally agreed world of public knowledge shared by observers of the system, whereas entities on the left-hand side exist in world 1 (the world of matter) (Popper 1979). Second, although Rosen advances the depiction as a way of looking at the formation of scientific theory, he argues at length (1987 and elsewhere) that the two sides of the picture are not equivalent: he believes that those systems for which there exist an encoding and decoding that capture the essence of the left-hand side reality are actually (very) special.

In terms of this type of picture, the standard approach or 'classical paradigm' (Malcolm *et al.* 1989) should look like that shown in figure 2.

The agent, on the left, contains a model of its environment and of its own capabilities, maintained in correspondence with the state of the physical environment and agent by the processes of perception and action. Acting on this model by a process of inference is the agent's controller. (Notice that we have drawn the agent enclosed by the environment.) The formal model of the system on the right comprises a model of the environment and agent and a model of the agent's controlling inference process. However, the classical paradigm asserts that the latter is in fact identical to the controller itself: on the right it is interpreted by the semantic machinery in the observers' heads whereas on the left it must be animated in some other way. For the environment and agent models and the controller, the encoding and decoding functions are the identity (or more rigorously an isomorphism). In effect, the observer's model of the agent is written in its brain for all to see.

In terms of this picture, the physical symbol system hypothesis amounts to saying that an encoding and decoding exist which make the diagram in figure 2 commute. Although the decoding might be expected to be straightforward, since

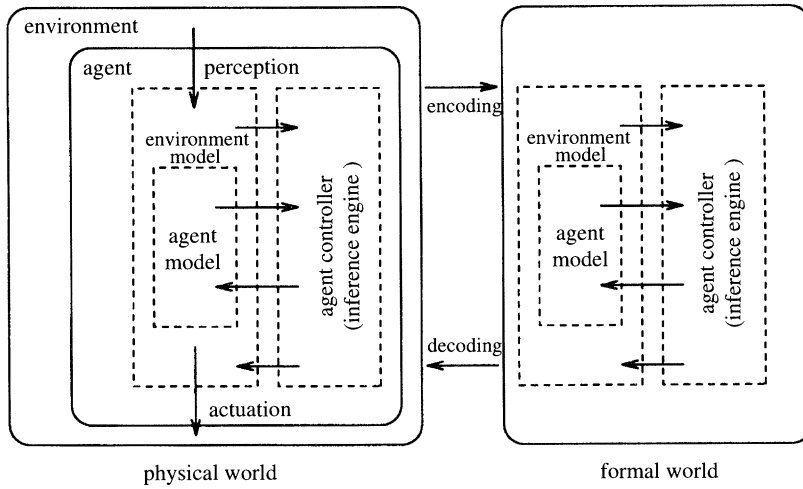

Figure 2. Comparing classical and situated-embodied paradigms.

any formal description possesses a decoding, the encoding is not; Harnad (1990) argues that the encoding is so difficult as to provide the dominating constraints on the design of the whole system, whereas, if Rosen is right, the encoding may not in fact exist.

This approach to the control of an agent is, in theory at least, perfectly feasible. It does, however, impose some strong constraints on the agent, since the internal structure of the agent and the formal model describing it must be isomorphic. These constraints in particular impact on the perception and actuation processes. A number of further remarks are in order.

First, notice that the formal description of the agent (agent, environment and controller models) can be said to have meaning in two distinct ways. From the point of view of the observer (i.e. 'meaning' from the observer's point of view) it acquires a derived meaning because it is expressed in the observer's formal language: a language whose meaning is presumably established by consensus among the observers watching the agent. This derived meaning in the observer's terms, however, does not necessarily imply that the formal description is meaningful in the same way to the robot itself. For this to be case the formal representation must play the same part in causing the robot's behaviour as it does in the explanation of the robot's behaviour offered by the observer. For very limited domains, and very limited tasks, this is apparently easy to achieve, because using the explanatory model in the implementation of the robot's behaviour will guarantee that the behaviour follows the model. This apparent ease is, however, an illusion, since all the hard problems of making sure that the robot's internal model corresponds with its circumstances have been solved for it by the observer: the observer chooses the right circumstances and switches the robot on. For robots with longer 'lifespans', intended to cope with varying circumstances, the robot itself must solve the problems of maintaining conformity between its model and its circumstances. Making this happen is an aspect of the symbol grounding problem (Harnad 1990). We agree with Harnad's claim that the difficulty of solving this problem is often greatly underestimated.

Second, the transparency of the cognitive architecture of the agent enables the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

legitimate use of introspection. If the agent is capable of introspecting, it does so by examining an isomorphic formal process to the right-hand side description, so any record of the observed formal process will be a valid right-hand side of the picture and can be used to explain and predict the activity of the agent.

Finally, with such a view of the agent it is tempting to work entirely on the right of the picture. After all, the formal models on the right are isomorphic to those inside the agent on the left, so investigation of them as formal systems must throw light on the cognitive structure of the agent. Essentially, this is what happens in classical planning and knowledge representation: assumptions are made about the content of the environment, and agent models and procedures for their manipulation are investigated, the hope being that perceptual and actuational systems consistent with those assumptions can be realized. That task, of course, falls on the engineers (or, if not, those who work on vision)!

Unfortunately, such a view of cognition makes life very hard for those interested in autonomous robots, for they have to build agents after the pattern of figure 2 and they do not have the luxury of delegating the perceptual and actuational encoding to other (often hypothetical) people. They are brought face to face with two critical problems: the dynamic nature of reality and the imprecision of sensing and actuation.

The latter problem impacts principally on the nature of the formal model the agent carries in its head. It must be possible to combine partly inconsistent measurements from a variety of sensory systems to construct a coherent formal view on the basis of which actions may be planned. For numerically quantifiable uncertainty, some progress has been made using the tools of bayesian statistics and the ubiquitous Kalman filter but the general question of describing the uncertainty endemic in the world in a formally well-founded manner is still open.

The former problem has two manifestations. The formal models used must be able to describe action and change in a coherent way using the syntactic tools of logic, and there are technical problems in doing so. Also, the agent must ultimately make decisions in a time consonant with the natural dynamics of the environment in which it is placed. This means that the manipulations of the formal models, which constitute the operation of the controller, must take place fast enough for the agent to keep abreast of its environment. In view of the representational complexity necessary to deal with the dynamics of a world sufficiently complex to be interesting, and the number of objects and relations such a world might comprise, the operation of the agent in real time is computationally intractable, at least with present technology. Furthermore, at any moment most of the representation, so costly to maintain, will be irrelevant as far as the agent's immediate decision is concerned.

Thus, for the roboticist, the classical paradigm makes the realization of an autonomous agent acting in a reasonably complex world an intellectually and computationally intractable problem. Although it can be made to work for simple static environments, all interesting real environments are dynamic and (usually) complex.

3. Situatedness and embodiment

Those who nevertheless pursue the implementation of robots using the classical knowledge-based architecture take the view that this current intractability is

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

simply one of those very hard problems which are very hard because that's the way things are. It will get easier when computer technology delivers us more powerful machinery and tools or we gain a deeper understanding of symbolic knowledge representation. We are concerned in this paper with those who take the view that the intractability of this approach (with today's technology) is due to taking the wrong approach to the implementation of intelligence, not to the nature of things.

Over the last ten years, some people in the robotics and planning communities have devised an alternative paradigm with a view to avoiding this artefactual intractability, and the results of this endeavour go under the name of situated robotics, behaviour-based robotics, or reactive planning and situated activity (in parts of the planning community).

This new approach rests on two independent presumptions about intelligent systems: they must be situated and embodied. We also identify embeddedness as a (rather stronger) alternative to situatedness as the latter is currently used in the literature. We shall discuss these in turn.

Situatedness, as the weakest of these presumptions, merits first treatment. It is used in the literature as a technical term to refer to those cases where an agent's controller interacts directly with the environment of the agent rather than in a manner mediated by an internal formal description of that environment. In Brooks's language, 'the world is its own best model' (Brooks 1991), or as Chapman (1990) puts it, 'an agent's most important resource in computing what to do is its current situation'. This view, taken seriously, has the immediate consequence that the agent must be able to recover from its environment by appropriate sensing whatever it needs to know to determine its course of action, and thus must be able to participate in its environment. In turn, this implies that the internal dynamics of the agent operate at a speed consonant with those of the external world.

A good example of a situated system is Chapman's *Sonja* (1990). *Sonja* plays a video game called *Amazon*, in which an explorer is faced with various challenges in a dynamic, partly visible, rapidly changing world. The system watches the game display, as a human would do in a game arcade, tries to interpret the visual cues therein, and selects from a number of possible actions at each possible moment those (or the one) most appropriate for the current situation, and performs the action chosen. Seen from within, *Sonja* has sensors (its visual system), actuators (the commands it can give to the game), and is constrained to play *Amazon* at the natural speed of the game.

Instead of using an 'objective' description of the environment with which it interacts and having, therefore, to simulate the dynamics of the world of *Amazon*, *Sonja* uses a deictic representation in which items are designated in terms of their current relation to the agent. The instantiation of such relationships is perceptually a much simpler task than for objective representations, and crucially, does not require keeping track of object identity. *Sonja* is also able to take linguistic advice on what actions and goals it should consider. Chapman advances the system as a model of routine activity and advice taking.

Sonja is interesting as it illustrates one end of the spectrum of situated agents. The world in which *Sonja* is situated is completely simulated; though the visual perception of the system is intended to mimic realistically the perceptual tools used by human players. Thus, it exists entirely on the right-hand side of Rosen's

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

picture. However, if we projected it on the left-hand side by building an agent that watched a real video game and operated its controls, the difference from an agent built under the classical approach would be the absence of a formal descriptive system inside the agent.

In a sense the classical paradigm, with which we contrast the situated approach, is at the other end of the spectrum. The work of Rosenschein & Kaelbling (1986) on situated automata illustrates the range of possibilities available. In their work, propositional knowledge on the right-hand side of the Rosen diagram is represented in the agent by state bits whose value is objectively correlated with the truth of those propositions. Inferences drawn on the right-hand side are compiled into fixed-depth combinatorial circuitry implementing the necessary state changes. In principle, one could compile the interpreter of the reasoning system which executes the agent description on the right into the circuitry of the agent on the left, to a greater or lesser extent. A full compilation (if possible) would result in something like the classical architecture, whereas a minimal compilation results in an agent with no ability to manipulate its goal structure at run time.

Stronger than situatedness is the requirement that an intelligent agent be *embedded* in its world. Embeddedness will here be taken as a technical term to refer to those cases where important aspects of the mentality of an agent are immanent in the interaction between the agent and its local world, and cannot be discovered solely in the architecture or states of the agent. For example, intrinsically relational states such as purposes make no sense except within embedded systems.

This idea may also be found in Brooks (1991), where two maxims are distilled from a discussion of intelligence and emergence, namely 'intelligence is determined by the dynamics of the interaction with the world' and 'intelligence is in the eye of the observer', because it is an emergent phenomenon arising from the interactions of many components, no one of which can be pointed to as the seat of intelligence. A trivial example of this idea might be a wall-following agent that achieves its task by moving in a shallow curve and 'bouncing' off the wall at regular intervals using a touch sensor. The behaviour of the agent arises because of the interaction with the environment; the competence is not local to the agent itself.

Embeddedness implies participation in the environment, i.e. situatedness, but is rather stronger since it asserts that the mentality of the agent is not delimited by the boundaries of its controller, across which information is transduced by perception and actuation, nor can it necessarily be delimited by the conventional boundaries of what we as external observers would refer to as 'the agent'. Rather, as Bateson (1972) pointed out,

> we must take into account the system – that is, the network of *closed* circuits, ... whose boundaries do not at all coincide with the boundaries either of the body or of what is popularly called the 'self'. (Bateson's emphasis.)

(We consider that *Sonja* is in fact an embedded system in the sense above, though it is not described as such by Chapman; deictic representations, in virtue of their dependence on the interrelationship between agent and objects, transcend the agent's boundaries. Thus *Sonja*'s intelligence is not solely a result of the internal manipulations it performs. We suspect that Brooks would disagree with this suggestion.)

The final presumption to consider is *embodiment*, here taken as a technical

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

term to refer to a physically embodied agent which interacts directly with the world by means of physical sensors and physical actuators, i.e. by means of direct physical causation. Thus a robot is embodied, whereas an expert system is not.

There are several reasons one can advance for the importance of embodiment. First, and most practical, is the point that the designer of an embodied agent must deal with reality on its own terms: unrealistic simplifying assumptions about the sensory or actuational problems faced by the agent are untenable. The agent must also interact with the environment at a rate determined by the latter, since we do not (yet?) have the technology to interrupt the causal processes of the real world.

Second, an embodied agent can in fact exploit its interaction with the world to simplify the perceptual and motor problems it would otherwise face. Wehner (1987) contends that animal sensory systems do this often, in that they comprise 'matched filters': in other words they exploit the physics of their sensory systems to enhance their ability to discriminate perceptually important stimuli. As an example one could consider the cricket whose 'ears' are connected by a tracheal tube one quarter of a wavelength long at the frequency of the cricket's song. This tube makes each ear of the pair strongly direction-sensitive and contributes to the cricket's ability to localize sounds. The human pinna serves a similar function by a different mechanism. Curiously, the cat extends its low frequency hearing by utilizing the tuned resonance of its *optic* nerve, whose vibrations it can sense, to register sounds up to an octave below those detectable by its ears.

Embodiment also introduces contingency into agents. No two agents can experience the world in precisely the same way because they cannot simultaneously occupy the same space, even supposing that we could build them identical bodies using different matter. One consequence of this is that it becomes necessary to study robots, as embodied agents, using the techniques developed by biologists for studying animals: every robot is unavoidably an individual.

Finally, an embodied agent engages in the causal pathways of which the real world is made, and this physical engagement with reality provides the ultimate grounding for any semantic machinery inside the agent. Without such physical grounding, any semantic entities found in the agent exist only in the eye of the beholder and are completely devoid of intrinsic meaning.

Notice, however, that embodiment is (at least in theory) independent of situatedness and embeddedness. The classical cognitive architecture discussed above is embodied, provided we really do build the agent shown on the left-hand side of figure 2, while the *Sonja* system is situated but disembodied.

It is also apparent that mere embodiment is not sufficient: there are different modes of embodiment to consider. Computer software, like any object in Popper's third world, can be embodied by recording it in an agreed formal language; however, software can also be embodied by enacting it with a suitable embodied interpreter. (Note that execution by a Turing machine or other abstract machine does not constitute embodiment in our sense.) Physical grounding entails embodiment in the latter sense.

To summarize, for those who suspect that at least some aspects of intelligence are not to be found in the neurophysiological complexities of the brain, or any 'software' the brain may host, but reside (at least partly) in the complex interactive commerce between the intelligent creature and its local world, designing, implementing and studying an *embodied* mind *embedded* in its local world is the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

only kind of AI research possible. This is why in recent years a number of AI researchers have moved away from the 'brain-in-a-vat' GOFAI computer-based AI, and begun experimenting with various kinds of robots; and why robotics, for so long the engineering-tainted foundling of AI research, with its own separate conferences and journals, has recently started re-appearing in mainstream AI journals and conferences.

4. Sensing, action and behaviour

The requirement that an agent be embedded in its environment has a number of interesting consequences. In this section we shall explore them a little.

Consider a simple mobile robot built to follow walls. As suggested above, the robot could be equipped with a wall-contact sensor on the right and a tendency to veer toward the right as it moves forward. (It veers right so that, by moving in a curved path, it will eventually hit a wall close enough to it on the right.) By moving forward until it hits the wall, then turning left by a suitable amount, this simple agent will successfully follow a wall on its right by 'bouncing' off it.

Such an agent is clearly embedded in its environment. There is no state bit in the agent that corresponds to the presence of the wall nor to the fact that a wall-following behaviour is being executed. In the sense of Rosenschein & Kaelbling it is not possible to say the agent *knows* that a wall is present or that it is following one. (Interestingly, if the agent can observe state trajectories over time then it is possible for it to know that it is wall-following, but this transcends the knowledge-as-objective-correlation-of-state view of the situated automata approach.) Nevertheless, observed wall-following behaviour in such an agent is quite reliable, and this competence essentially presupposes the tactile interaction with the environment that we have sketched.

Even for this simple agent, it is not sufficient to give the putative program the agent is executing. 'Move forward, turn left $x°$ if contact is detected' is not a sufficient prescription for wall-following: some of its 'mind' is outside the agent. In particular, the robot does not move forward when commanded to; rather it veers right. Even incorporating this into the program statement is not enough. Without (external) knowledge of the presence of the wall, the behaviour generated by the program is indeterminate.

The fact that, even for a simple case such as this, the agent's behaviour is not fully determined by its program leads one to wonder. Of course, one could argue that the mobile robot described above is rather poorly engineered, and allowing it to veer to the right under program control would make it more versatile. A short step further,then, is to suggest that, if the robot were correctly engineered, the program would be a sufficient description of its behaviour. Experience suggests, however, that the kind of issue raised by the veering robot is fundamental: the assumptions made by the engineer about the behaviour of the agent do not always correspond with the actual behaviour in reality. To permit abstraction of interaction into program in this way requires perfectly understood and predictable sensing and actuation, and this is (at the very least) a hard engineering problem.

What is really going on when we try to use the program as a specification for the mobile robot in the example above is that we are breaking the causal couplings responsible for the embeddedness of the agent. The behaviour of wall-following is

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

realized by the causal coupling of processes inside the (conventionally delimited) agent and processes in the environment. In choosing to focus on the program, we have cut the causal connection out in the environment and supposed that the robot is an input–output transducer. The environment generates states that are transduced by sensors into program inputs; the program transforms inputs to outputs; and outputs are transduced by actuators into effects in the environment. This dissection, of course, is called the computational metaphor; it allows us to view the processes inside the agent algorithmically.

This is not the only possible way of cutting the causal loop. Rather than cut the loop in the environment, we might cut it in the agent. This is the perspective of perceptual control theory: the agent generates actions to try to achieve particular sensory states (Powers 1973). For example, a mobile robot attempting to pass through a door might first try to align itself so that the edges of the door frame seen in its visual sensor were symmetrically placed about the centre of its field of view. With the agent–world loop cut like this, it appears that the environment is performing transduction between actions of the agent and the agent's perceptions.

It is of course possible to analyse a closed loop system, such as an embedded agent, in this way, by cutting the loop somewhere and inspecting the open loop system thus created. However, something is lost in doing so.

As a final observation, notice that whichever cut is taken, perception and action appear only after dissection. Perception transduces across the boundary from environment into the agent, whereas action (or actuation) transduces across the boundary in the opposite direction. However, these categories or processes have been introduced by the dissection: they are, in fact, artefactual. For example, by using 'sensitive actions' (effector mechanisms with stable behavioural outcomes selected by environmental parameters), it is possible to construct a wall-follower system which behaves externally exactly like the mobile robot described above but which contains no components that would be recognized as sensors if removed from the agent (see, for example, Malcolm 1995).

Similar devices that can robustly explore the floor of a room are available in toy shops for a few pounds as toys for pets. They consist of a ball containing a motor, battery and asymmetrically mounted weight, attached by gearing to a diametric spindle. When the motor is driven, it attempts to lift the weight but instead the ball rolls around the spindle and moves forward. If the ball is obstructed, however, the weight is swung upward by the motor, making the ball unstable. It moves to restore equilibrium, the weight falling to the bottom again and the ball rolling off in a new direction.

This kind of 'sensitive action' also shows that sensing and actuation are not natural kinds, but products of a certain kind of dissection usually, but not always, possible. The conclusion, then, is that for embedded systems at least, it is behaviour that is fundamental, while perception and actuation are conventional and often convenient categories for describing the function of the system, but should be used with caution: they already presuppose a dissected system.

5. Purpose, semantics, and embeddedness

Just as there are two kinds of semantics which something may have, there are two kinds of purpose. The two kinds of semantics are *ascribed* and *intrinsic*. Ascribed, derived, or attributed semantics is the kind we as readers ascribe to

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

books. Intrinsic semantics is possessed by the properly grounded symbol system of an autonomous robot. The important difference is that intrinsic semantics involves the autonomous operation of an *active* symbol system and its connection to its external referents via the causal links used in its interaction with the world.

Similarly we ascribe purposes to tools such as pencils or tin-openers. The purpose derives from the use we make of them, and the pursuit of the purpose requires our direct immediate purposeful and causal involvement with the device. We can also build autonomously purposeful devices, such as the thermostat. It is true that we designed and use the thermostat for our own (intrinsic) purposes, but the important difference is that we delegate the achievement of the purpose to the device. It is, in its own minimal way, an autonomous agent which tries to achieve the purpose by means of its own self-directed behaviour, and can continue to do so after its creators and users (and their purposes) have ceased to exist.

Some people may wish to argue that, as far as purposes are concerned, the most important difference is between biological agents (the only things which can have *real* purposes) and everything else, from tools to artificial 'agents', all of whose purposes are ultimately derived from the purposes of biological agents. Others will argue that only human beings have real purposes, from which all other purposes derive. We do not wish to argue that there are not important distinctions to be drawn between the purposes of robots, worms, dogs and human beings. We simply wish to draw attention to the distinction we have described here between ascribed and intrinsic purpose, because this has particular importance when considering the various architectures and kinds of mentality which artificial creatures may have. To assert the importance of this distinction in this context is not to deny the importance of other distinctions in other contexts.

Following Malcolm's suggestion we shall refer to as *autoteleology* the kind of intrinsic purposefulness displayed by autonomously operating goal-seeking devices such as servomechanisms. The simplest possible kind of autoteleological device is a homeostatic system such as a system comprising a room, heater, power source and thermostat. It is important not to forget that when we say 'a thermostat has the purpose of stabilizing the temperature' we are using the convenient synecdoche of referring to the entire system by naming a characteristic component. A thermostat on its own is useless, a fragment of a system, and can no more have a purpose than a brain in a bottle. Autoteleology can only be had by systems which have interactive commerce with the local world in which they are embedded. A goal-seeking system, such as a servo-mechanism, involves an active causal loop cycling between the mechanism and its local world. As we have explained in the previous section, for analytical purposes we can break this loop at various points in the cycle to give various different viewpoints on the system, but we must not forget that terms such as 'purpose' apply only to the entire system, and strictly speaking cannot be applied to particular components dissected out of the system.

For these reasons neither a programmed computer nor a brain-in-a-bottle can, strictly speaking, be said to have purposes, although we often say so, employing the shorthand synecdoche of using these distinctive components to refer to the whole system. If, however, we make the mistake of supposing that a programmed computer could, on its own, host a purpose, and therefore embark upon a research programme to discover just how to do this, we may find the problem strangely intractable.

This is what Searle suggests by his Chinese room argument (Searle 1980),

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

and why Harnad (1989) suggests that robots (of appropriate architecture) are immune to this criticism. It is a certainly a step in the right direction to move from the computer (or brain) to the robot (or creature), but our contention is that a further step is necessary: many mentalistic terms, such as purpose, can only properly be ascribed to systems comprising *both* the agent *and* the world in which it goes about its business.

Of course, Searle's argument is not about purposes, but understanding and semantics. It is no coincidence, however, that we find the same two kinds of semantics as we do of purposes: ascribed and intrinsic. Let us refer to intrinsic semantics as *autosemantics* (Malcolm 1995). It is easier to see the importance of embeddedness to autoteleology, because it is easy to see that goal-seeking devices require interactive causal loops between the device and its environment. Meaning is a more abstract concept than purpose. An autoteleological device can be constructed entirely in terms of physically causal effects, such as Watt's steam-engine governor, or a thermostat (using these characteristic components of the entire systems to stand for the whole). Autosemantics, however, requires further architectural steps: to begin with, the system must be an information processing system.

It has been argued that a thermostatic system is the minimal symbolic system, containing one or two symbols, in terms of which it expresses beliefs, such as 'it is too hot'. Others have argued that, for the terminology appropriate to symbolic systems (belief, error, truth, etc.) to apply, the symbolic system must have a number of symbols, a syntax, inference machinery, and so on. How many symbols? How much syntax? Our argument depends on no specific answers to these questions, merely that at some stage of complexity we have a symbolic system to which it is appropriate to ascribe beliefs etc.

It is characteristic of all the robots so far constructed by roboticists, and of all the animals so far analysed by biologists, that at the lower levels of sensorimotor interaction with the world, great use is made of goal-seeking mechanisms, i.e. of control. In autonomous agents, whether biological or artificial, the purposes and parameters of these goal-seeking mechanisms can provide a very convenient ready-made source of grounded symbols†. While there must also be other mechanisms involved (e.g. to cover the case of an autonomous agent which has the purpose of finding an algorithm for generating the nth prime number)‡ there is no doubt about the importance of these goal-seeking mechanisms as a source of symbols, because they can neatly encapsulate both many of the capabilities of the agent, and the perceptions relevant to these capabilities. Malcolm & Smithers (1990) and Hallam (1994) have built robot systems illustrating the hosting (grounding) of GOFAI symbolic systems in this way, using goal-seeking mechanisms, and claim that this is a particularly simple and computationally efficient architecture, avoiding the intractabilities characteristic of other methods of grounding symbol systems. Symbols which are grounded via autoteleological mechanisms will naturally thereby have intrinsic meaning for the agent so hosting them, i.e.

† Although frequent use is made of the term 'grounded symbol', in a grounded symbol system it is not just the symbols which have to grounded, but the entire system, syntax and inference mechanisms too, in order to achieve the necessary isomorphism with the world and its dynamics.

‡ A much sought but theoretically unfindable entity, which, even if it did exist, would have no material existence.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

autosemantics, whereas symbols grounded via derived purposes, such as in an expert medical diagnosis system (grounded via the expertise and sensorimotor capabilities of the people using the system) will be ungrounded.

Of course not all symbols need be grounded in this way. Intrinsically relational or broad content symbols, as philosophers call them, are wider in scope than autoteleological derivation. They, nevertheless, have the same crucial property that their description cannot be entirely in terms of the internal states of the agent, but must involve some reference to the agent's local world and its traffic with it. So even if they are not grounded by the kind of autoteleological mechanisms we have been discussing, they are clearly grounded in a similar way by whatever mechanisms subserve their referential and functional relationships with the rest of the symbol system and with the world.

Thus the argument in favour of embedded systems, which we have presented largely in terms of purposes and autoteleology, has a higher-level and more complex parallel in terms of meaning and autosemantics. In other words, trying to find autosemantics in the central nervous system of animals, or trying to implement autosemantics in a computer, is a doomed enterprise, because autosemantics, like autoteleology, is a property only of the complete historical and contingent agent–world interactive system.

Hence the importance of embedded robotics to AI research.

6. Conclusions

In summary, we have seen that an approach to the understanding of cognition is possible which eschews formal representations as necessary components of the causal engine of an agent and sees cognition as necessarily embodied and embedded: grounded in physical reality and extending beyond the confines of the conventionally delimited agent, to include those aspects of its local world with which the agent interacts. Rosen's argument suggests that, even were we to take the 'agent' to mean the system in an extended sense, still a purely formal representational account would fail to capture the essence of cognition. In the long run, therefore, we will only succeed in building highly versatile cognitive agents when we have a good idea what sort of virtual machinery goes inside their heads, and we can only discover that by building embodied and embedded agents whose semantics are grounded (primarily) autoteleologically.

We thank Bridget Hallam, Gillian Hayes and Alan Bundy for comments on the draft of this paper, and the University of Edinburgh for its provision of computing and text preparation facilities.

References

Bateson, G. 1971 The cybernetics of 'self': a theory of alcoholism. *Psychiatry* **34**, 1–18.

Brooks, R. A. 1991 Intelligence without reason. In *Proc. 12th IJCAI* (ed. J. Mylopoulos & R. Reiter), pp. 569–595.

Chapman, D. 1991 *Vision, instruction and action.* Cambridge, Mass.: MIT Press.

Dreyfus, H. L. 1979 *What computers can't do : the limits of artificial intelligence.* Cambridge, Mass.: MIT Press.

Hallam, J. 1983 Resolving observer motion by object tracking. In *Proc. 8th IJCAI*, pp. 792–798.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Hallam, J. 1994 Playing with toy cars : an experiment in real-time control. DAI Research Paper No. 527.

Harnad, S. 1989 Minds, machines, and Searle. *J. exper. theor. Artificial Intelligence* **1**.

Harnad, S. 1990 The symbol grounding problem. *Physica* D**42**, 335–346.

Haugeland, J. 1985 *Artificial intelligence: the very idea.* Bradford Books, MIT Press.

Malcolm, C. 1995 *Behaviour, purpose and meaning.* In preparation.

Malcolm, C. & Smithers, T. 1990 Symbol grounding via a hybrid architecture in an autonomous assembly system. In *Architectures for autonomous agents* (ed. P. Maes). MIT Press, North-Holland.

Malcolm, C., Smithers, T. & Hallam, J. 1989 An emerging paradigm in robot architecture. In *Intelligent autonomous systems 2* (ed. T. Kanade, F.C.O. Groen & L.O. Hertzberger). Amsterdam.

Newell, A. & Simon, H. 1981 Computer science as empirical enquiry. In *Mind design* (ed. J. Haugeland). Bradford Books, MIT Press.

Popper, K. 1979 *Objective knowledge.* Oxford: Clarendon Press.

Powers, W. T. 1973 *Behaviour: the control of perception.* Wildwood House.

Rosen, R. 1987 On the scope of syntactics in mathematics and science: the machine metaphor. In *Real brains, artificial minds* (ed. J. L. Casti & A. Karlqvist). Elsevier.

Rosenschein, S. J. & Kaelbling, L. P. 1986 The synthesis of machines with provable epistemic properties. In *Proc. Conf. on Theoretical Aspects of Reasoning about Knowledge* (ed. J. Halpern), pp. 83–98. Morgan Kaufmann.

Searle, J. 1980 Minds, brains, and programs. *Behavioral Brain Sciences* **3**, 417–457.

Wehner, R. 1987 'Matched filters' – neural models of the external world. *J. Comp. Physiol.* A**161**, 511–531.

Discussion

D. Partridge (*University of Exeter, U.K.*). Dr Hallam has stressed that in robotics, solutions are constrained by the available hardware. Why should we expect the robotics explanations to apply to biological hardware?

J. C. T. Hallam. Building autonomous robots won't give us a general theory of intelligence – but nor will classical AI. A general theory may not be attainable, only specific solutions that aren't hardware independent.

M. Brady (*University of Oxford, U.K.*). There are two extreme approaches. One stresses abstract computational accounts, the other biological 'wetware'. Robotics is an intermediate stopping point, which forces us to confront the issues faced by biological intelligence.

R. Hudson (*University College London, U.K.*). Learning and memory presumably play a part in robotics? Speaking as a real human, I find I often don't react to the world as it really is, but as I remember it being. So I bump into things that weren't there before, and avoid things that are no longer there. This suboptimal performance should be the standard for robot behaviour if we want robotics to give insights into how our minds work.