From COGWHEELS to COGNITION

The Knowledgeable Robot

Chris Malcolm

Institute of Perception Action & Behaviour,
Division of Informatics *,
Edinburgh University,
5 Forrest Hill,
Edinburgh EH1 2QL,
Scotland, UK
Chris.Malcolm@ed.ac.uk

Abstract. Recent progress in the contributing disciplines of cognitive science (philosophy of mind, neurophysiology, genetics, ethology, evolutionary theory, computer science, artificial intelligence, etc.) have made it possible and interesting to compare the "mental" capabilities of people, animals, parts of minds, and computer systems. Whereas computer scientists routinely use metaphorically extended terms like "decide" to refer to decision-like processes in computers, some philosophers of mind insist that such mentalistic terms be restricted to the operations of the full orchestra of the human mind. Consequently terminological confusion has confounded the interesting debate about the extent to which machines could ever be made to think using this or that approach. This paper proposes ways of developing more terminologies with which we can talk more clearly about the varieties of mental capabilities available to different kinds of creatures and devices, both natural and artificial. Focussing on the questions of attributing knowledge to a robot, and how knowledge may be implemented in a robot, leads to the development of the idea of seeing the robot mind/body complex as a hierarchy of technology domains. This in turn leads into a cybernetic or behaviour-based view of mind, encompassing the interactive complex of brain, body, and world, in which brain (or computer) is not the host of mind, but its generative organ.

Stent [Scientific American Dec 1972] also considers the possibility that an entire field of enquiry may lack connexion ... with the body of knowledge possessed by the scientific community at large. This is precisely the situation of machine intelligence research, which uses the techniques of computing to investigate questions not of natural science but of philosophy. [Michie 73]

Technology is at present covert philosophy; the point is to make it openly philosophical.[Agre 97], p240.

1 INTRODUCTION

"Artificial intelligence" is an ambiguous phrase. It could mean an artificial substitute for real intelligence, something which in the absence of the real thing can provide a useful enough approximation. A computer which can play chess is in that sense an existing example of artificial intelligence, and no more controversial than an artificial milk substitute.

A more controversial interpretation of "artificial intelligence" is real intelligence which has been constructed artificially. This is the stated long-term goal of some of those working in artificial intelligence, and naturally carries the implication that it is at least debatable that giving a machine some kind of mind is in principle possible. This has stirred much controversy, since there are many ways in which mind can be held to be special in a way which would forbid its implementation in a computer, from arguing that current computers have the wrong sort of architecture, through the idea that biological brains have some crucial properties impossible for electronics to capture, to the idea that human minds are special in a way beyond the reach of science.

A lot of this controversy is due to the poor and ambiguous vocabulary we have for discussing mentality. In particular there is a wide gap between the meanings of common mentalistic terms such as "understand", "think", "decide" as used by those with philosophical backgrounds, and those with computational backgrounds. It is the purpose of this paper to develop terms for some of these mental and quasi-mental things which clarify these important differences of meaning, and facilitate their discussion. Developing a taxonomy of these shades of meaning

* Includes what was, before 1999, the Department of Artificial Intelligence

for mentalistic terms, and pursuing their application to the questions of why robots may be said to know things, and the ways in which these kinds of knowing may be implemented, turns out to lead into a justification for a behaviour-based or cybernetic view of mind.

The argument starts by introducing the computational metaphor for mind to show that there need be no dualistic mystery about the way in which mind (or knowledge) affects matter, and introduces robotics as the synthetic approach to investigating creaturehood, just as AI is the synthetic approach to investigating problem solving. It proceeds to identify three different strengths of the concept of understanding based on criteria for identifying the presence of understanding. The next question to consider is where we should expect to find understanding implemented in animals, and where we should be implementing it in robots. The classical answer to both questions is "in the computational mind", and the assumptions behind this view are clarified. Assuming this for the sake of argument, a taxonomy of the various different kinds of knowledge that could be implemented in the computational mind (or computer) is developed, one kind of which is *dispersed* knowledge, often called *emergent* or *distributed*. This kind of knowledge develops considerably more complexity when we turn from considering computers to considering robots, because interacting with the world via sensors and actuators is considerably more complex than interacting with people via text input (keyboard) and text output (screen). The idea of a structured hierarchy of technology domains is introduced to make sense of this complexity. Given the costliness of brains to any creature which must carry them around with self contained power supplies, this naturally leads to the behaviour-based approach to robotics, and a cybernetic view of knowledge and mind which spans the interactive complex of the behaviour of a creature in its local world. As far as robots and animals are concerned, this spoils the separability hypotheses on which the computational view of mind was based. The argument concludes by showing that there are reasons for supposing that this behaviour-based or cybernetic model of mind has promise of developing further, rather than ending up in a specialised niche like insects or expert systems.

2 The COMPUTATIONAL METAPHOR

Computer software shares the weightless and dimensionless characteristics of mind which led Descartes to assign mind a separate ontological status. Yet despite this apparently immaterial nature, it has an understandable two-way causal traffic with the physical world via the physical machinery of the computer. We do not have to suppose new laws of nature in order to explain how computer programs can cause physical changes in the world. It is instructive to consider why not.

In Aristotelean hylomorphic terms software is the form for which computer hardware provides the substance. To exist at all software must be encoded somehow, but the medium doesn't matter, and in most cases the writing or erasure of software involves no change in weight or substance, just in arrangement. Nor is any particular encoding *the* program; a program can only be stopped from existing by erasing *all* copies in all media, including those from which it could be reconstituted, such as listings in books. Yet, like an idea, it can be stolen by copying it. Consequently we regard software as essentially an abstract thing, a substanceless form.

Having decided on a particular means of encoding, however, Shannon's theorem allows us to calculate the minimum amount of substance required to encode (write) a certain piece of software [Shannon and Weaver 49]. Although only the merest fragment of matter is required to encode a symbol, *some* fragment is required, and without some such fragment somewhere, it does not exist. That fragment is enough to affect the relevant symbol reading machinery, which in turn affects the symbol interpretation machinery, and so on. In other words, tiny though the material part of a symbol (or piece of software) is, it is the physical causal effect of that fragment which leads, by a chain of further physical causation, to the ultimate required physical action in the world. In fact the economical virtues of small size in the elements of information processing machinery will tend to drive these elements to the furthest extreme of miniaturisation which the interpretation machinery can reliably support. This extremity of smallness, and the independence of encoded symbols of any particular kind of material manifestation, has led some to overlook their nevertheless *essential* materiality, essential because it gives the capability to *cause* effects in the material world.

Symbols, or computer programs, immaterial though they are in one sense, nevertheless affect the physical world by the ordinary causal powers of their residual but essential materiality.

Thus computer software and hardware give us an example of a dualism which resembles Cartesian mind/body dualism, but in which the causal traffic between the two requires no magic, no new laws of nature, nothing we do not currently well understand. A computer program is nevertheless a very different kind of thing from an ordinarily material thing such as a stone, or the hardware of a computer. So there is a strong ontological dualism between programs and stones, but not quite strong enough to forbid causal traffic completely. In fact it is an essential feature of programs that they are associated with very specific kinds of physical machinery which mediate their causal traffic with the world.

I do not wish to claim the strong form of the computational metaphor which asserts that minds are like computer software and brains are like computer hardware. What this discussion of informed machinery demonstrates is that, unlikely as it seemed a few hundred years ago, there is a way in which something with most of the properties of Cartesian Spirit can causally interact with Cartesian Matter. The missing property is consciousness. In other words we know how to make zombies (non-conscious robots). Can zombies be further developed into conscious beings? Are worms zombies? Are bacteria?

These are interesting questions which are being given increasing amounts of attention. One of the things which makes this particularly difficult is that our vocabulary for discussing information and mentation is vague and ambiguous when faced with these new challenges. Another is the complexity and messiness of biology. Even a worm is a formidably complex machine with levels of functional organisation extending at least into the realm of individual chemical molecules, and we don't have any good ideas how to find out if worms are conscious in Nagel's sense of "is it like something to be a worm?" [Nagel 74], i.e., is there is a subjective experiential quality to being a worm?

Where cognitive psychology studies problem solving analytically, by observing and experimenting with problem solving animals, artificial intelligence studies problem solving synthetically, by making artificial problem solvers. Making them permits us to pose questions of principle and architecture very clearly, by embodying them in the design. Problem solving, however, is only one aspect of the deeper business of being a purposeful knowledgeable autonomous agent, which I shall refer to as being a creature, since that word nicely accommodates both animals (God's creatures) and robots (our creatures). A particularly difficult aspect is what makes the difference between knowledge and nonsense, which involves such difficult notions as semantics, intentionality, symbol grounding, and purposeful agency.

This is where robots come in. Just as artificial intelligence offers us a way of studying problem solving synthetically, so robotics allows us to study what is involved in creaturehood by making artificial creatures. This paper uses what we know about knowledgeable robots to develop our ideas of knowledge. Knowledge manifests itself in knowledgeable behaviour, behaviour which we consider implies understanding. The most direct way of testing for its existence is by means of tests, examinations, etc.. The issue is unfortunately complicated by the existence of cheating and various other ways of appearing to understand while not really understanding. This poses a problem for the builders of robots and other systems which they want to claim do in fact have some kind of understanding: how is the claim to be tested?

That contentious question has a long history in debates between Artificial Intelligence and its critics, so I shall begin by considering what we mean by understanding.

3 VARIETIES of UNDERSTANDING

It is one of the methods in the tool kit of philosophy to distinguish different versions of the concepts involved in an argument, and to see how varying the concepts affects the argument. The most common kind of variation is between strong versions of a concept, which involve more assumptions, and weaker versions, which involve fewer assumptions. Here I will distinguish between three strengths of the idea of understanding.

UnderstandingA (**A***lan Turing Understanding*):- This is the weakest variety, and is defined purely in behavioural terms: if the candidate understander produces the right behaviour then it understands. This will be known as *Alan Turing Understanding,* or **understandingA** after the Turing Test [Turing 50].

UnderstandingB (**B***rian Smith Understanding*):- A memorious parrot[1] could succeed in demonstrating **understandingA**. Many people don't like the idea of attributing understanding to memorious parrots. This problem afflicts all behavioural tests of understanding. Brian Smith gave us the following compact characterisation of a stronger form of understanding, in his *Knowledge Representation Hypothesis*:–

> *Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but causal and essential role in engendering the behaviour that manifests that knowledge.* [Smith 82].

Here, not only is the right behaviour important, as in **understandingA**, but it must also be produced by the right kind of machinery. Smith intended this as a characterisation of the hypothesis, usually unstated, which underlay research in the major AI paradigm, variously known as logicism, cognitivism, symbolic AI, knowledge-based

[1] Often technically referred to as the Giant Lookup Table of all possible questions and answers.

systems, and GOFAI (Good Old-Fashioned AI, a term coined by Haugeland in [Haugeland 85])[2]. Smith proposed a particular mechanism, namely propositional representation and inference, but I use it here to stand for the more general idea that the *right kind of mechanism* must be involved in understanding, without being specific about what the right mechanism might be.

UnderstandingC (**C***onscious Understanding*):- This kind of understanding, typically human, and essentially involving our subjective experience of understanding, will be called *Conscious Understanding* or **understandingC**. It is the kind of understanding that (Searle claims) a Chinese person has and the Chinese Room has not [Searle 80]: understanding *involving* the conscious experience of understanding. I use the word *involving* to indicate a strong version of conscious understanding, in which consciousness plays an *essential* causal role in the process of understanding. There is a weaker epiphenomenal version which claims that consciousness is involved with understanding only for accidental reasons of the way we happened to evolve [Humphrey 92].

In sum we have defined three kinds of understanding:

UnderstandingA, **A***lan Turing Understanding*, behaviourally defined;
UnderstandingB, **B***rian Smith Understanding*, producing the behaviour by the right kind of machinery; and
UnderstandingC, **C***onscious Understanding*, involving the subjective experience of understanding.

Note, by the way, that I do not claim that understanding has only these three varieties. As I indicated in the discussion of **understandingC**, this can be further divided into epiphenomenal and causally functional consciousness. One could also distinguish between fully reflective role-playing consciousness and simple unreflective awareness. This tripartite **-A, -B, -C** division is just a useful tool for initial exploration.

3.1 ESCAPING from SEARLE'S CHINESE ROOM

My purpose in this section is to exemplify the utility of distinguishing these varieties of understanding by showing that they facilitate escape from the apparent Schank/Searle Chinese Room problem.

Schank asserted that a computer system he was developing already showed the beginnings of real understanding [Schank and Abelson 77]. It used English language input and output, and given some clues about an instance of a commonplace scenario (such as eating in a restaurant and not paying the bill), used a highly structured knowledge base to draw "common sense" conclusions (such as something being wrong with the food or service). Searle reacted to this by supposing in a thought experiment that he (Searle) pretended to be the computer executing a successfully "understanding" Schankian computer program, and demonstrated by analysing this thought experiment that **understandingC** was clearly not present [Searle 80]. This is hardly surprising, since no attempt to implement **understandingC** had been made, merely an attempt to implement **understandingB**.

Seen in terms of these different varieties of understanding, a great deal of the fire & smoke of the Chinese Room debate disperses. There is a sense in which we can accept both Searle's Chinese Room Argument and Schank's assertion without contradiction — Schank claimed he had made a step towards a certain goal, **understandingB**, and implied that this was a step towards **understandingC**. Searle retorted that repeated elaborations of **understandingB** wouldn't ever get him to **understandingC**. The subsequent debate has been long and confusing [Hofstadter & Dennett 82]. The interesting question of whether **understandingB** might be an important component of **understandingC** (e.g. necessary but insufficient) has often lain unnoticed in the scuffle.

4 MIND, BODY, and WORLD

Having developed these three varieties of the concept of understanding in terms of the criteria for identifying its existence, I next wish to turn to the question of where we should expect to find it implemented. "In brains" is the common and obvious answer, which leads to the natural corollary "and the computers controlling robots". I wish to unpick and refine these, because I later wish to cast some doubt on them.

We observe intelligent behaviour in a creature going about its business in its local world. The behaviour is something the creature does. We also observe that some creatures characteristically display more intelligent behaviour than others. We naturally conclude that the key to intelligent behaviour is to be found in the creature rather than its circumstances.

Knowing from studying damaged creatures that the brain is responsible for coördinating behaviour, connecting sensory inputs to motor outputs, and so on, we naturally conclude that the key to intelligent behaviour is to be found in the brain.

[2] As often happens is science, Smith crafted this carefully worded definition of the hypothesis because he wished to cast doubt on it, but most of those who had been following this research programme had supposed it so obvious that they hadn't bothered to make their assumptions explicit.

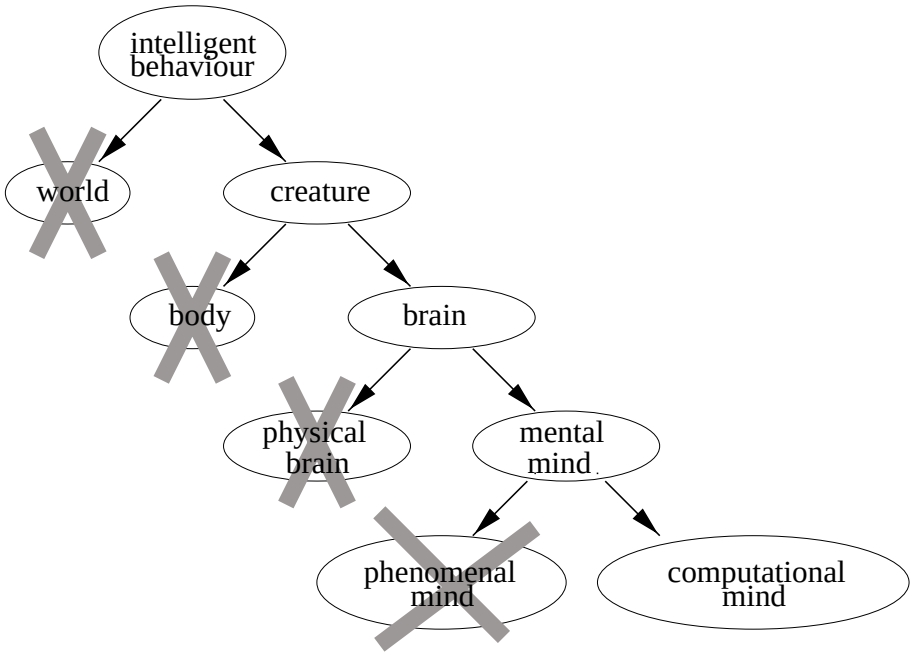

Fig. 1. Searching for the cause of intelligent behaviour.

Many models of mind share the very general assumption that there is a level of description of mental function which is independent of the particular technology supporting it[3]. In the knowledge-based systems research programme it is asserted that this level of description consists of the knowledge known and the access and inference machinery which can use it. Newell and Simon are articulate exponents of this view [Newell & Simon 76,Newell 81]. The underlying computational machinery can be made of transistors or clockwork, it doesn't matter. Thus (for as long as the analogy holds) it doesn't matter whether the underlying machinery is made of meat (a brain) or silicon (a computer) — if the knowledge and the functionality of its machinery are the same, then the behaviour will be the same. In short, you could (if you knew how) implement a real mind in a computer. In effect this is a claim that at this level of description of mind a virtual machine (a virtual mind) exists. It does not of course need to claim that this is the only level of description at which a virtual mind exists.

So, returning to the main thread of the argument, if, as the symbolic functionalist model of mind asserts, mentality shares the functional independence characteristic of computers and programs, then we can distinguish mind from brain, and decide that the secret of intelligent behaviour is to be found in the mind, rather than the (physical) brain.

Hobbes and Leibniz suspected long ago that it was possible to separate the computational mind and the phenomenal mind. The phenomenal mind has feelings, experiences, sensations, etc., whereas the computational mind follows chains of reasoning, associations, analogies, and so on. It is an assumption of the knowledge-based paradigm in artificial intelligence that the computational mind can be separated from the phenomenal mind, and that the secret of intelligent behaviour is to be found in the computational mind [Pylyshyn 86].

Note that there are two aspects to the computational mind in the human being. Most obviously there is the deliberate rule following we perform by consciously attending to the rule-following process, as in doing long division by hand. As we know, this can be automated in a non-conscious machine[4]. There are also nonconscious processes, those which deliver ready-made fully-fledged conclusions, hunches, and perceptions to conscious awareness. It is a moot question whether these are implemented in the brain as rule-following processes. What we do know, as a result of AI implementations, is that at least some of them can be successfully implemented in computers in this way. It is therefore possible, at least to some extent, to build artificial minds with human-like problem solving abilities, without regard to the functions of consciousness and feelings in the human being, and without bothering

[3] It is a crucial feature of the modern computer that functional description is independent of technology.

[4] This is no accident. Computers were designed in the image of a mathematician pretending to be a rule-following machine in order to perform the publicly formalised steps of a mathematical procedure.

where in the human being the lines are drawn between conscious thought, subconscious but potentially conscious thought, and ineluctably unconscious thought[5].

Getting from the starting point of wondering what causes intelligent behaviour to isolating the quarry in this kind of pure symbolic knowledge and inference has taken a number of steps. To recapitulate, we separated the behaviour into creature and world, separated the creature into body and brain, separated the brain into physiological brain and mental mind, and separated the mind into computational and phenomenal aspects (see figure 1 on page 5). In fact it is only required that *computational* mind is separable from brain; *phenomenal* mind could be entangled inextricably in brain or bodily physiology without upsetting the basis of the knowledge-based paradigm.

There is no doubt that there is *some* truth in this knowledge-based systems hypothesis, since successful problem-solving machines have been built on this basis, and the fountain of research and development based on these assumptions, although subject to much critical attack, is far from drying up. The question raised by the critical attacks, however, is whether this is the whole story of intelligent behaviour, or whether there is more to it, perhaps much more.

The classical knowledge-based view in Artificial Intelligence, characterised by Brian Smith's *Knowledge Representation Hypothesis*, is that intelligence is knowledgeable behaviour, and the crucial issues in implementing it are deciding what needs to be known, and how to represent and use it. Computer programs represent the "computational mind" of computers or robots, and are used in automated thought experiments (i.e. computer programs) to simulate theories about the "computational mind" of human beings and animals. In the sixty or so years of making computers and computer-controlled machinery (such as robots) do clever things we have learned quite a lot about this, and we now routinely teach computer science students guidelines about good and bad ways of encoding knowledge in programs.

The next steps in this argument are to develop some simple ideas about knowledge representation in computers, and then to extend them to robots.

5 VARIETIES of KNOWLEDGE

Previously we distinguished three varieties of understanding based on identification criteria, types **-A, -B, & -C** (see 3).

A similar tripartite distinction can be made of kinds of knowledge and kinds of meaning. In this section we will be concerned specifically with **knowledgeA&B**, **meaningA&B**, etc., in other words, not the full conscious human forms of these mentalistic concepts, but those which can, with our current understanding, be implemented in computer systems and robots.

5.1 EXPLICIT, IMPLICIT, and PROCEDURAL KNOWLEDGE

We have already considered one variety of **knowledgeB** — the symbolic knowledge of knowledge-based systems. There are two aspects to a knowledge-based system: the knowledge base; and the access and inference machinery which uses it, and which can add to it by making deductions. There are correspondingly three types of knowledge involved here: *explicit knowledge*, which is encoded explicitly; *implicit knowledge*, which can be derived from the explicit knowledge (and thus made explicit, if desired) by the inference machinery; and the procedurally encoded know-how of the inferencing. In human terms, explicit knowledge is what I know directly, and implicit knowledge is what I could deduce if I could be bothered. These are both types of symbolic knowledge. The important feature of symbolic knowledge is that because it is expressed (or expressible) symbolically, it can be reasoned about — and can in principle be combined with — *any* other item of symbolic knowledge in some chain of reasoning.

Computer scientists refer to this kind of symbolic knowledge as declarative, to distinguish it from procedural or algorithmic encoding. Although some kinds of knowledge fall naturally into one category rather than the other, it is usually possible to translate from one to the other. Where the knowledge is entirely specific to a certain skill or method, and there is no requirement to reason about it, it is often more economical of computer memory and time to encode it procedurally.

We shall call this kind of procedural knowledge **procedural**.

[5] Is unconscious thought *really* thought? It is part of the argument of this paper that we should stretch our mentalistic concepts in this kind of way, while paying attention to the kind of stretching we are doing. Otherwise we will end up in the kind of tedious verbosities suffered by "scientific" descriptions of animal behaviour which try to exclude purpose.

5.2 DISPERSED KNOWLEDGE

Procedural knowledge is non-symbolic, but easily identifiable as residing in a particular part of the whole system. There is a further kind of knowledge which is neither symbolic nor easily locatable, namely *dispersed knowledge*. This kind of know-how emerges from the interaction of a number of components of the system. It is built in to the design of the entire system, but there is no particular place where it can be identified. For example, a chess program may behave as though it "knows" that commanding the centre of the board is important, and its designer may have deliberately built it that way, but the centre-commanding behaviour may emerge from the interaction between the weightings given to several kinds of move, a few fragments of scattered heuristic code affecting pawn placings, and the natural property of the chess board that more moves intersect in the centre.

Some call this kind of knowledge *emergent*, some *distributed*. I prefer the term *dispersed*, since it conveys both the fact that it is distributed or scattered, plus the implication that it might be hard to identify all the pieces. It also avoids associating it with the technical computer science use of "distributed", and the philosophically contentious term "emergence". So, we shall call this kind of knowledge, which is built in to the system and its relationship with its local world in a various and dispersed manner, **dispersed**.

Following Haugeland in [Haugeland 98] both of these kinds of non-symbolic knowledge, **procedural** and **dispersed**, we shall refer to as **tacit** because, not being symbolic, they can't be reasoned about, they can't be "spoken" about in the creature's "language of thought"[6].

Thus within the computer system we have the two broad categories of **symbolic** and **tacit** knowledge, the **symbolic** knowledge divided into **explicit** and **implicit** kinds, and the **tacit** knowledge divided into the **procedural** and **dispersed** kinds.

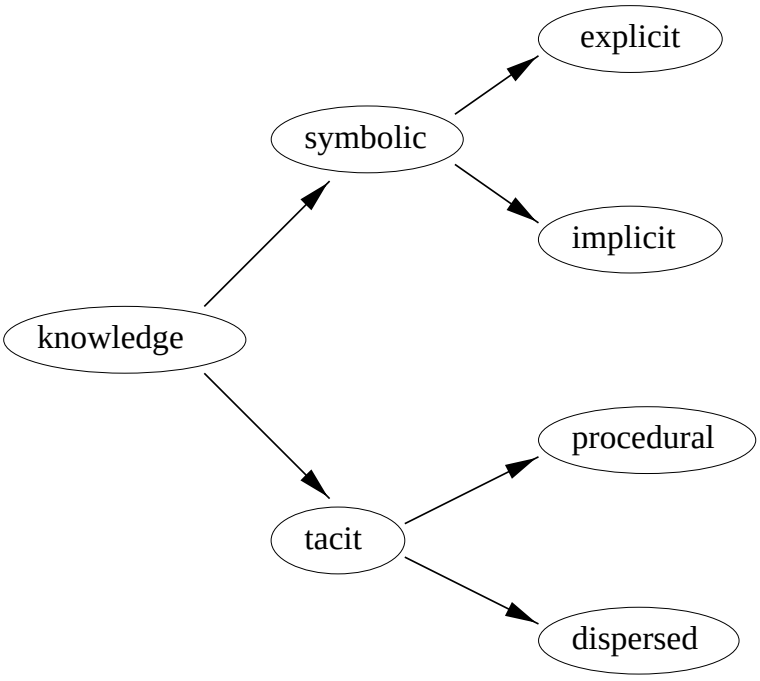

Fig. 2. A taxonomy of types of knowledge

Note that in categorising any kind of knowledge in this way it is necessary to specify the knower. For example, it is possible for a subsystem to have **explicit** knowledge which is not available to the system as a whole, so that the same knowledge which is **explicit** to the subsystem as knower is **tacit** to the main system. For example, suppose that a system is capable of roughly understanding and responding to English questions in English (as are some current database enquiry systems). This system could have a large linguistic subsystem which encoded some approximate rules of English grammar explicitly as some kind of augmented production rules. Yet these

[6] Although the term "tacit knowledge" has a long pedigree in philosophy of mind, Haugeland uses it here quite specifically in an AI context and in the **knowledgeA&B** senses I use it here.

rules might be kept privately within the language subsystem, and not be available to the system as a whole. Thus this English grammar would be **explicit** knowledge from the point of view of the language subsystem, and **tacit** knowledge from the point of view of the system as a whole.

This is a simple example of an important general concept: the attributes of knowledge are relative to a knower, and may change as the knower changes, even if one knower is part of the other.

So far we have considered knowledge in computers. Things get more complex when we consider robots.

6 LEVELS of IMPLEMENTATION

It is a commonplace of computer science that if one wishes to implement some particular function, such as finding the prime factors of a number, that it can be done at a variety of levels in the machine:–

- It can be programmed directly in a high level programming language;
- it can be provided as a function in the high level language by having been encoded in the low level language in terms of which the high level language is written;
- it can be provided as a function in the low level language by having been provided in the machine microcode;
- and finally, it could be directly implemented in the logical hardware which hosts the basic machine functions.

In sum it can be provided in high level software, low level software, firmware, or hardware. It can be provided at all these levels of computation because each level is a virtual machine with full capabilities. Of course this is a simplified picture. In practice here are often more levels, the implementation of a particular function may be split across more than one level, and virtual machines are sometimes incomplete, necessitating some unprincipled and unmodular bodging to provide certain functions.

6.1 The IGNORANCE of SPRINGS and LEVERS

In robots there are even more levels. A classic example from the world of assembly robotics is Nevins & Whitney's *Remote Centre Compliance* [Nevins and Whitney 78] (see diagram on 9).

They intended to solve the problem of getting a robot to insert a peg into a hole without jamming, despite misalignment, by using a mathematical model of the physics of the peg-in-hole problem to deduce the adjusting motions from the sensed reactive forces produced by the misalignment. During their investigations, however, they discovered that an arrangement of springs and levers in the gripper could be devised which would adjust the motion of the gripped peg in exactly the required fashion — by causing rotation about a virtual remote centre of compliance. Hence the name of the device they invented[7].

Now, had Nevins and Whitney succeeded in their original aim, we would have been able to see a robot sensing the reactive forces and knowledgeably adjusting its motions so as to put misaligned pegs neatly into holes[8]. We would naturally suppose that the robot in some sense "knew" about pegs and holes, and used this knowledge to generate appropriate movements. And if we looked inside the robot's brain we would indeed find, as expected, a mathematical model encapsulating exactly that sort of knowledge, which was being used to produce the behaviour[9]. This situation exemplifies nicely both Newell & Simon's *Physical Symbol System Hypothesis* [Newell & Simon 76], and Brian Smith's *Knowledge Representation Hypothesis* [Smith 82].

So far so good. Now consider another similar robot using the Remote Centre Compliance, with its springs and levers, and let us suppose that the mechanism is concealed. We shall see this robot behaving just as did the previous one. We will again naturally conclude that this knowledgeable behaviour will be generated from some appropriate internally represented knowledge. But this time we won't find it. Instead we will find a physical mechanism doing the work. Is the knowledge-based robot behaving knowledgeably, but the identically behaving device-based robot not behaving knowledgeably?

[7] The remote centre of compliance is situated at the hole-entering (front) end of the peg, and is therefore remote from any compliant part of the mechanism, in particular being remote from the gripped (back) end of the peg, which normally provides a centre of compliance there, and when misaligned leads to jamming, which the RCC avoids. The intuition behind this device is the realisation that a peg pushed into a hole from the back end has a tendency to jam if misaligned, whereas if the peg is pulled into the hole from the front end, it has a tendency to avoid jamming and realign itself. The RCC makes a back-end-push behave like a front-end-pull.

[8] It is possible to do this, and many have done so, it is just slower and more expensive.

[9] In some implementations the mathematical model will be much more easily visible than others. It is always possible to arrange that it will be easily visible.

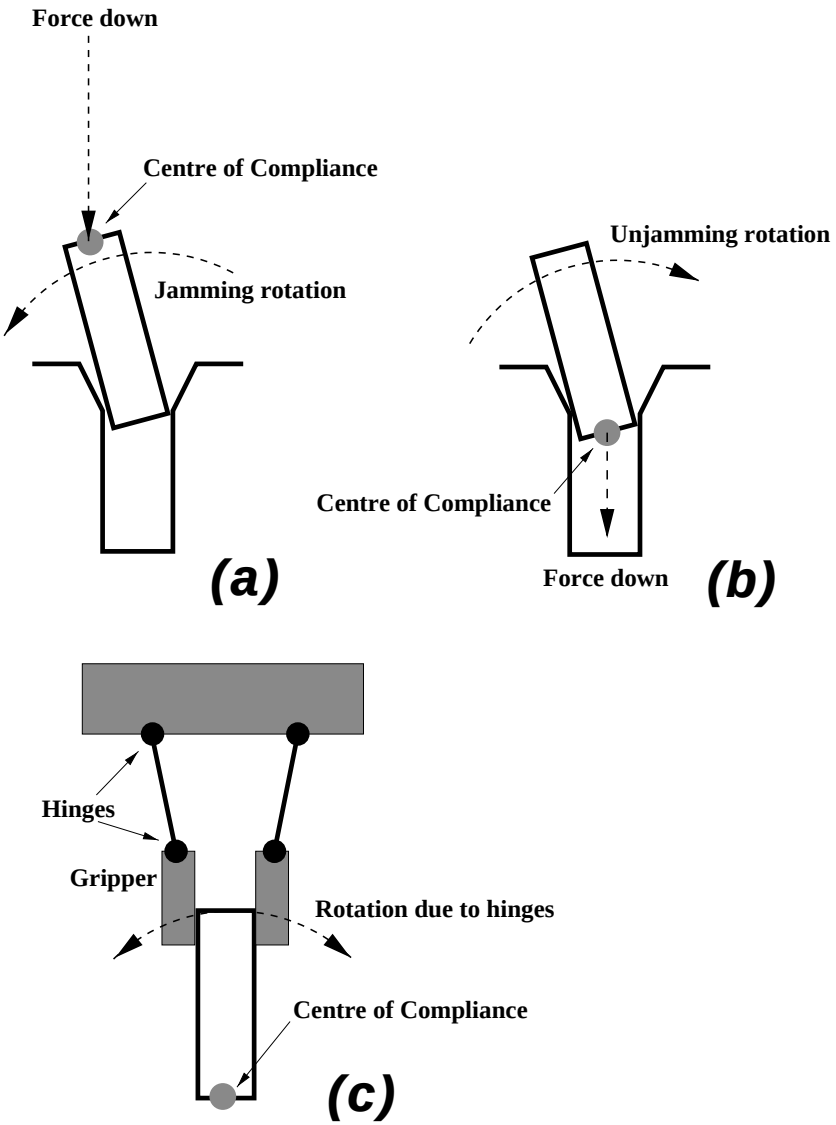

Fig. 3. The Remote Centre Compliance. In *(a)* the peg jams because the forces push the top end of the peg to the side. In (b) the peg slides in because the forces push the top end of the peg into alignment. In *(c)* the hinges arrange that the pushing forces from the top cause rotation about the bottom of the peg as in *(b*. [The actual device is more complex than this. This simplified diagram concentrates on the essential eponymous feature.]

As far as **knowledgeA** is concerned, a robot which puts a peg into a hole by consulting a mathematical model, and a robot which does it by means of the springs and levers of a Remote Centre Compliance, are equally knowledgeable because equally capable. Note, however, that **knowledgeB**, which permits scrutiny of the internal machinery for the required specific characteristics, would allow us to outlaw the springs and levers as ignorant. We could therefore describe the robot using the force sensor and mathematical model as doing peg-in-hole by means of **explicit knowledgeB**, whereas the second robot has changed the **explicit knowledgeB** to **tacit knowledgeA** by moving it from the computational to the mechanical level. The two categories of **tacit** knowledge we have so far are **procedural** and **dispersed**. It is not **dispersed** here because it is entirely located within one modular mechanical unit, but it is not **procedural** because it is not algorithmic. The RCC is in effect a mechanical analogue computer, but operating with the large causally effective energies themselves rather than the tiny signal level energies usually employed in analogue computers.

In short, we have here a kind of knowledge in a robot which is not implemented computationally in the robot's "brain", but as an appropriately behaving device in the physics of the robot's "body". It will be necessary to extend our categories of knowledge in order to accommodate this kind of thing. This involves developing some terminology for describing the many layered architectures of creatures' bodies. For example, the simple *sense* $\rightarrow$ *think* $\rightarrow$ *act* model of of robot control involves a division between the thinking layer, which is naturally handled in the AI or CS department, and the sensor/actuator layer, which is naturally handled in the electronic and mechanical engineering departments. We need to consider more layers than this simple model, because it turns out that these layers, and their relationships, have important parts to play in the story.

7 TECHNOLOGY DOMAINS

A technology domain consists of a kit of general purpose devices which exploit the device-making possibilities of some causal domain, plus some fastening and linking bits and pieces, and a general theory and methodology of how to make devices with this kit. It is also necessary to have transducers which permit causal interaction between this and other technology domains. A transducer is a device which transforms one kind of energy into another, such as electricity into motion, or sound into electricity. Thus transducers convert between the different forms of energy characteristic of different technology domains. They are the bridges between technology domains. All sensors and effectors are at bottom transducers. In fact, in the more abstract domains the devices which connect domains are not what physicists would call transducers, although they are analogous. I have adopted the term *transductor* for this kind of generalised transducer.

A good example in a toy realm of the mechanical domain is a child's Meccano or Lego Technic set. There are plates, struts, beams, fasteners, axles, gears, pulleys, string, chains and chain wheels. These are characteristic components of the mechanical technology domain. Both Meccano and Lego provide extension sets which extend the basic mechanical domain of the elementary sets into the electrical domain. Apart from elements of the purely electrical domain, such as wires, relays, switches, and batteries, it is also necessary to have transductors which permit causal interaction between the mechanical and electrical domain. Typical transductors here are electric motors, solenoids, micro switches, magnets and reed relays, and lights and light sensors. The electrical domain is characterised by the use of electric currents of appropriate size for causal interaction with the mechanical world. The handling of currents of this size, with the heat dissipation and forces involved, sets a strict lower bound on the size of these components. For example, an electric motor powerful enough to lift as much as you can, with similar speed, would be difficult for you to lift, and would need a fairly chunky switch to turn it on and off.

The next technology domain in this example hierarchy is the electronic domain. This domain is characterised by the use of very small electrical potentials, too small to be causally efficacious in the electrical domain, and which are essentially signals. The characteristic tools of this domain are differential amplifiers, oscillators, filters, and their raw components, transistors, resistors, inductors, and capacitors. There are advantages to making these parts small. The use of photolithographic manufacturing techniques has made some of them microscopically small (VLSI silicon chips). The transductor from the electronic to the electrical domain is typically the amplifier. Many transductors from the mechanical world, such as optical sensors and microphones, naturally provide electronic scale signals.

The next technology domain in this hierarchy is the domain of digital electronics. The typical components here are the packaged VLSI modules known as silicon chips, such as logical gates, flip-flops, adders, counters, selectors, registers, and clocks. The transductors between this domain and the electronics domain are analogue-to-digital converters (ADCs), and digital-to-analogue converters (DACs). Some transductors from the mechanical domain can be arranged to have digital outputs, and some electrical switches can be directly driven from digital signals.

7.1 The COMPUTATIONAL DOMAIN

In this sequence of technology domains we have now arrived at the boundary between hardware and software: the computational domain. Digital electronic signals are already in the right form, but in order to enter the specially organised regime of a particular computer the input and output signals are gated via ports. The logical function of a port is to introduce the signals to the clocked and regulated regime of a computer.

Up to this point each technology domain has been more abstract and more versatile than its predecessor, and each technological kit of parts could be regarded as a virtual machine. To recapitulate we have passed from the mechanical domain, via the electrical domain, the analogue electronic domain, the digital electronic domain, finally to the computational domain. The first transductors were pure energy transformers, i.e., were transducers. As we ascended the hierarchy the physical energy transformations lessened, and features of organisation increased, to the point where it is not clear whether we should regard ADCs as transducers or translators, and computer input ports are essentially immigration control authorities. Hence the general term *transductor* to cover all these domain bridging devices. The computer is a completely general purpose information processor in which the processing is performed by rules of operation collected into entities known as programs, and from this point on the levels of virtual machinery are all software. The physical function of the transductor entirely ceases to exist, and what performs the transduction function is simply levels of software organisation.

If we imagine a robotic creature using artificial intelligence to guide its behaviour, then on the input side we start with the physical technology domains operating by means of physical causation. It is from these that the sensors are constructed. As we proceed up the input (sensor) hierarchy we substitute signals for the raw energies that activated the sensors. These signals are as small as they conveniently can be, and represent measurements of features of the original energies. As we proceed further up the hierarchy these signals become transformed into the abstract high level symbols of perception. This is the process known to cognitive psychologists as "categorical perception" [Harnad 87]. When after due thought the creature decides what to do, the process unravels in reverse down the output (effector) hierarchy, with symbols becoming signals becoming electrical power becoming physical forces which act upon the world.

This process is illustrated in figure 4 on page 12. That is just an example hierarchy in some notional electrically-powered robot. There are other technology domains, such as pneumatics. Note too that while the particular kind of technology does to some extent predispose the domain to a particular relative position in the hierarchy of some creature in which it is used, it does not compel the position. For example, while there are good reasons in today's technology for implementing the computational domain on digital electronic technology, it is possible, as Babbage tried to demonstrate, to implement computers in clockwork. The position of any domain in the hierarchy of any particular creature will be determined by the functional role it occupies within the architecture of that creature.

In a creature all of whose actions are decided at the highest level (or the only level) of its brain, there is a single loop of information transfer, from sensing the world up the levels of the input hierarchy to the level at which it will decide what to do next, and then down the output hierarchy from symbolic command to physical forces which act upon the world, in a repeated cycle from world to perception to thought to action and back again to world. The input perception hierarchy is linked to the output action hierarchy at the top level. This is often known as the *sense* $\rightarrow$ *think* $\rightarrow$ *act* cycle. That cycle is too simple for two reasons. Firstly in practice there have to be a variety of cycles of different speeds depending on different levels or urgency. This is well known, and has for a long time been part of the theory of computer operating systems. Secondly, even within one such cycle, there are subcycles or short-circuits within the sensing and acting limbs of the cycle. This purely robotic (or biological) aspect is not as well known as it should be, probably because it is in effect an inter-disciplinary study.

Rather like learning, which early AI hoped was something you could add on afterwards after you had got the rest of it right, this kind of extra-computational short-circuiting of the *sense* $\rightarrow$ *think* $\rightarrow$ *act* cycle turns out to be an important basic feature of creature architecture rather than a useful accessory. The next section will develop these short-circuiting ideas with the help of the Remote Centre Compliance.

7.2 SHORT-CIRCUITING the HIERARCHY

What we have seen in the case of the Remote Center Compliance (see page 8) is the short-circuiting of this hierarchy by building a cross connection at a lower level, in this case the low level of the mechanical technology domain. Similar short circuits can be built across at all levels. Of course it is not always possible to do this kind of short circuiting, and there are important general capabilities (such as planning) which are only available at the highest levels. Nevertheless, where it *is* possible to build such a short circuit, it saves computation and improves speed of response. It therefore makes sense to analyse a system employing short circuits like this in terms of **knowledgeA**, because this permits us to consider the knowledge as being dispersed throughout the levels of the

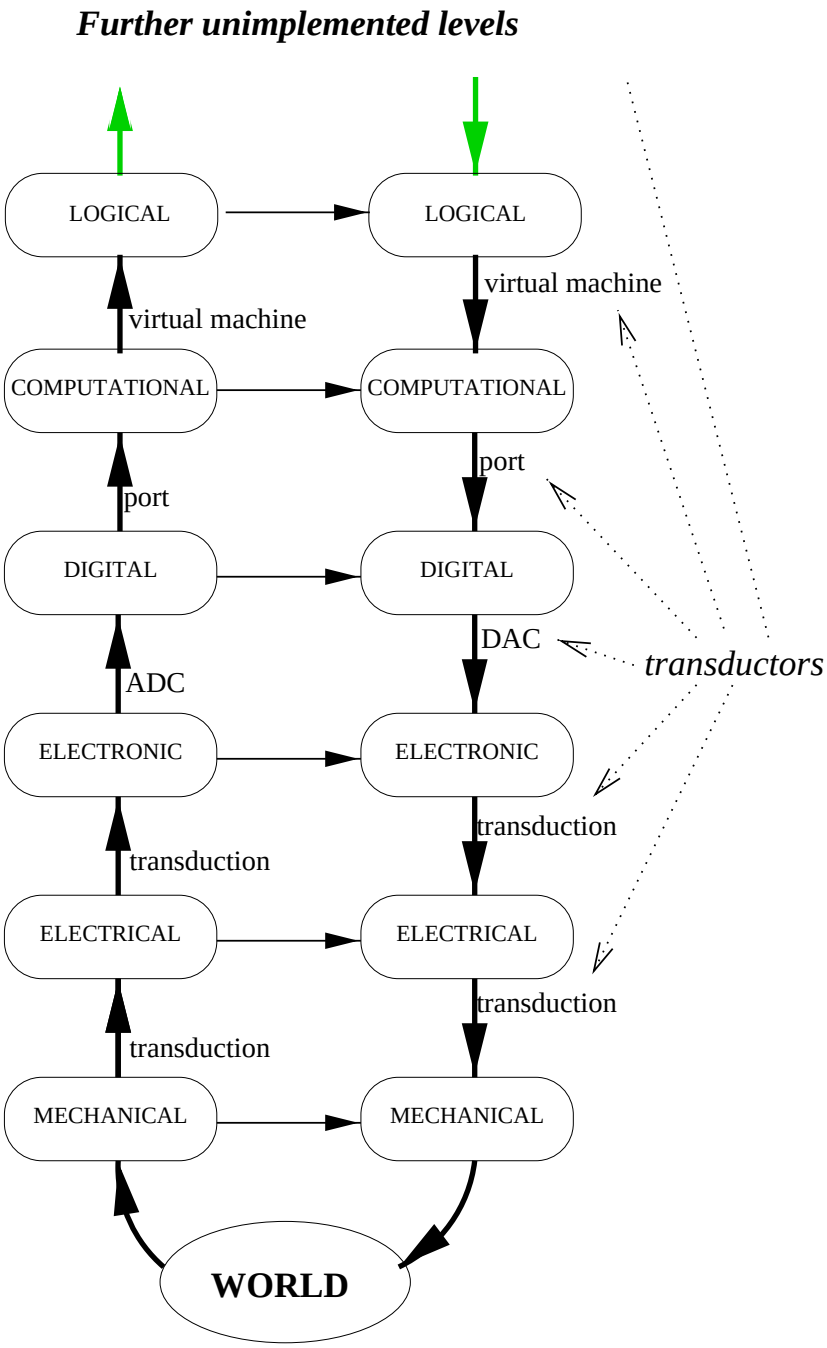

Fig. 4. Technology Domains

technology domains. Note that if we think in terms of **knowledgeB**, the knowledge simply disappears as it devolves down the hierarchy of technology domains, because **knowledgeB** is fussy about the way it is implemented.

This is not just a nice philosophical point. Robot designs sometimes come to grief because the design team delegates other parts of the technology domain hierarchy to others, and then ignores them. For example, at least one attempt to design a very fast powerful direct drive robot arm failed because of unforeseen interference between control exercised in the computational domain, and control exercised in the electronic domain. The amplifiers driving the motors utilised the standard technology of a power amplifier whose non-linearities were corrected by negative feedback of output error operating with a cycle time of roughly 1/50th of a millisecond. Normal robot arm joint position controllers operate with computer-based control loops of between 1 and 10 milliseconds, but this one was so high performance that its joint control loops were over a hundred times faster. It's an engineering rule of thumb that to stay out of trouble you operate a higher control loop at least ten times slower than a lower one. In this case the two control regimes interfered with one another with the result that the robot's motion was poorly controlled and sometimes unstable[10].

It took a lot of time and expense to identify the problem because the robot designers presumed the amplifiers could be considered perfect black boxes. So they could, so long as the resolution of their attention was worse then 1/5th of millisecond, just as the chair you're sitting on is solid only so long as the resolution of your attention is worse than atomic orbital speeds, and just as we regard plants as stationary creatures, but time-lapse photography shows them fighting each other for sunlight.

As these examples illustrate, timing adds a further possibility to the hierarchy of technology domains.

7.3 TIME AND DYNAMIC STABILITY

There are many aspects of robot construction which impose timing constraints.

- If (as is likely) some levels contain control functions, then this forces a slow-down on higher levels to avoid destructive control interference.
- It takes time to do certain necessary information processing.
- The physical dynamics of particular kinds of behaviour imposes strict constraints.

For example, as a tall roboticist who made large robots was fond of explaining, big animals fall over more slowly and therefore don't have to think so quickly, permitting the use of cheaper brains[11]. I need to react within months to calls for papers, within hours to hunger, within seconds when walking the streets, and so fast when treading on a sharp spike that evolution has arranged to do this reaction in the lower spinal cord rather than wait to communicate with the distant brain. By the time "I"[12] feel the pain of the spike, my foot is already on its way up.

Another architectural trick to improve timings, where different kinds of demand are made on one mode of sensing, is to develop different hierarchies, each specialised to one mode of operation. This has, for example, been done in the case of primate vision. One hierarchy, which is fast, largely confined to 2D image-plane computations, and focuses in differences, is used for the fast visual control of motion, such as ducking away from a blow aimed at your head. The other, which is slow, concerned with 3D layout in space, and concentrates on invariants, is used for object and free-space recognition, and the control of behaviour. Only the slow 3D hierarchy communicates with consciousness, hence the strange phenomenon of "blind sight", in which someone who is completely blind nevertheless shows certain kinds of visual awareness, without being conscious of them, when tested [Clark 99].

Once you have divided up a creature control system into layers with different timings in this way, a further architectural possibility presents itself. Lower levels can create dynamic stabilities by means of negative feedback of errors which higher levels (with their coarser temporal resolution) can take for granted as simply stable. For example, most of us take for granted our vertical stability when standing and walking, with no idea that this involves (among other things) visual feedback loops based on handy local architectural verticals. That is why climbing the Leaning Tower of Pisa is so weirdly unsettling.

This means that a higher level in the input (sensory) hierarchy can be supported not just by lower levels in the same hierarchy, but by dynamic stabilities arising from lower level sensorimotor "short circuits".

The Remote Centre Compliance of Nevins and Whitney (see page 9) gives a good example of how this hierarchical short circuitry can create a novel amalgam. I earlier presented this device as a kind of analogue computer, using causally effective forces, which connects sensor to actuator at the mechanical level. But where is the sensor?

[10] Unfortunately the design team neglected to publish this instructive failure.

[11] The late Jack Todd of Mechanical Engineering, Edinburgh University.

[12] Using "I" here in the narrow sense of my conscious centre of narrative gravity.

It is impossible to identify any part of the peg/hole/gripper/RCC complex as a sensor. Where is the actuator? It is impossible to identify any part as an actuator. In effect the entire complex is both sensor and actuator.

There are many such devices. I call them *sensitive actuators* because sensors and actuators cannot be distinguished within the recurrent loop of force translations which comprises the entire complex. Sensors and actuators are conceptual abstractions which do not apply in all cases of adaptive behaviour. This kind of sensitive action only occurs at the bottom of the domain hierarchy, but there are analogous things at higher levels, such as multi-stable oscillators.

Thus the introduction of timing has turned the original technology domain ladder diagram into more of a twisted rope ladder or double helix.

The classical views of creature architecture (e.g. the *sense* $\rightarrow$ *think* $\rightarrow$ *act* model) which ignore these kinds of things ignores important possibilities. Pursuing this topic is beyond the scope of this paper. Introducing it makes clear that our normal sensor/motor categories need extension and development as much as do our categories of knowledge.

Biological evolution is ruthlessly economical. Animals are only provided with features which have, in their evolutionary history, earned their keep. Our human brains are large and heavy, and on average consume about 20% of our energy. That's an expensive thing to carry around. No doubt having twice as much brains would be an advantage, but it would have to be advantageous enough to pay for the extra costs of carrying and running it before evolution would buy it. Much more attractive to evolution are ways of being smarter with the same size of brain.

Robot designers are involved in a much earlier example of the same compromise. If a robot is going to carry its own brain and batteries around, then its computational brain will be strictly limited in power and memory. With today's technology and methodology, the major problem of building a self-sufficient robot capable of crossing a busy multi-lane urban road while paying due attention to drunks, children, dogs, footballs, white sticks, traffic signals, etc., is keeping the whole thing small enough that crossing the road would be a physically feasible operation[13]. To get the most out of limited resources robot designers, like evolution, must exploit every possible trick for minimising computation and miniaturising the computer. Minimising computation involves not just computational tricks, but using technology and behaviour strategies which reduce the need for computation. Hence the importance of technology domains. In effect some of the computation can be implemented, or the need for it removed, in technology domains *below* the computational.

The research programme known as "behaviour-based robotics" (Varela refers to it as "enactive systems" [Varela 91]) is concerned to exploit this kind of dispersion of knowledge as much as possible, and has demonstrated that considerable computational savings can be made. Note that making these savings was one of its main motivations [Brooks 86,Malcolm et al 89,Pfeifer 96].

8 DISPERSAL OF KNOWLEDGE INTO THE WORLD

A lot of attention has recently been given to the business of dispersing knowledge beyond the confines of the creature into the world, e.g., [Clark 98], so I will simply mention it here to complete the picture.

At the top end of the brain scale we augment our own brainpower by the use of pencil and paper, lists, diagrams, maps, calculations, diaries, address books, reference manuals, textbooks, and these in turn are tools in the sociohistorical scientific method, a kind of socialised human learning whose power and capacity far exceeds that of any individual human mind or lifetime. We also employ this kind of **dispersed** knowledge in organisations. For example, there are several companies in the world who know how to make large long-distance passenger aeroplanes, but no single person does, nor does any collection of people without computers, reference books, and lots of pencils & paper. As literate creatures we are clearly *dispersing* **knowledgeB** & **knowledgeC** between people and into books etc.. At the other end of the social scale termites exploit social and environmental interaction to produce the design of their elaborate termitaries without any plan in any termite's head. Here what is **dispersed** between creatures and into the environment (via pheromones etc.) is **knowledgeA**.

Having now finished the development of our concepts of knowledge, we are in a position to review the original classical computational model of mind which found the locus of implementation of understanding to be the computational mind.

[13] Assuming, to make the task properly difficult, that it is legal to cross the road anywhere (as it is in Britain), not just at controlled crossings.

9 The SEPARABILITY HYPOTHESES

Recall the original hunt for the locus of the magic ingredient suspected to be at the root of intelligent behaviour. Starting from observing a creature going about its business in its local world in an intelligent manner, we chased intelligence from there into creature, into brain, into mind, and finally into computational mind. Each of these divisions involved a separability hypothesis, the hypothesis that our quarry was not hiding athwart the division.

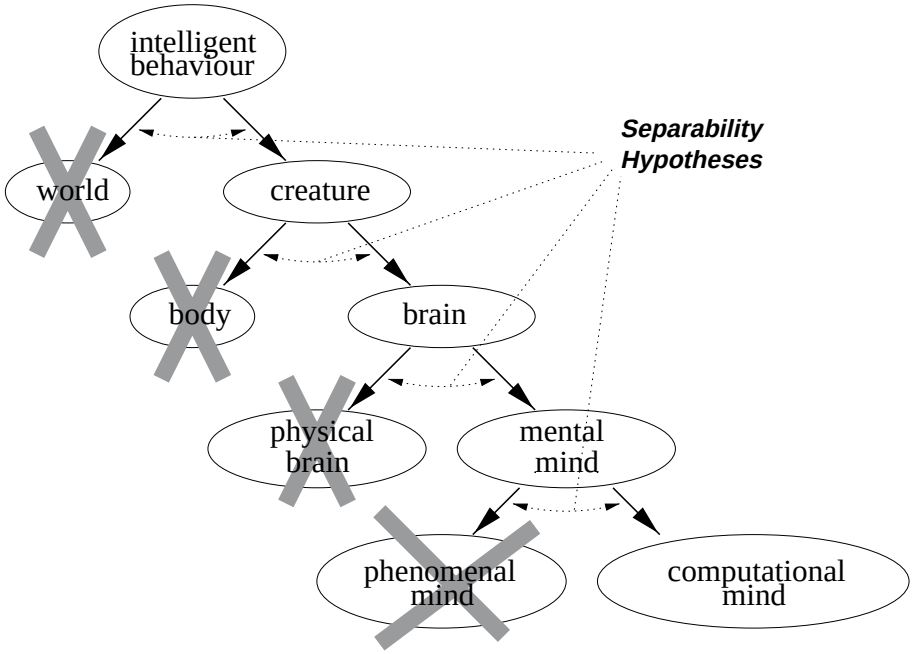

Fig. 5. Searching for the cause of intelligent behaviour.

The original presumption of the hunt was that intelligent behaviour was knowledgeable behaviour. Knowing now that some of the knowledge may be **dispersed** knowledge, residing in the design of the creature's body, and some of it may be **dispersed** into other creatures, and into the local world, emerging through interactions, we can see that none of these separability hypotheses hold securely.

It is unarguably very useful to have all relevant knowledge available simultaneously as explicit knowledge within one mind. Unfortunately, if a creature has to carry and feed its brain, in practice (both in biological evolution, and in robot design) it seems to be too expensive to be worth the computational and memory costs. As Canny of MIT says in Russell & Norvig's *Artificial Intelligence* [Russell and Norvig 95]:

> In robotics *the principal drawback of the classical* [knowledge-based] *view is that explicit reasoning about the effects of low level actions is too expensive to generate real-time behaviour.*[my insert]

In sum, since both biology and robotics find it a useful principle from the point of view of economy in creature design to exploit **tacit dispersed** knowledge as much as possible, then it follows that the separability hypotheses on which classical knowledge-based artificial intelligence is based cannot be maintained.

Of course, if one is concerned purely with computational implementations of intelligent computational assistance to people in some task, such as in the case of an expert medical diagnostic system, these arguments do not apply, since the system does not interact with the physical world, but interacts with people at a textual level, and the interacting with the world is done by the human users of the system. In fact these systems, as Winograd & Flores argue [Winograd & Flores 86], only display a mimicry of knowledgeable behaviour which is parasitic upon our knowledgeable use of them[14].

[14] They also argue that Artificial Intelligence can go no further than that, but they restrict their conception of Artificial Intelligence to the kind of classical knowledge-based systems to which that criticism applies.

10 WILL IT SCALE UP?

AI has long been haunted by wonderfully suggestive implementations of ingeniously simplified aspects of mentality which turned out not to scale up to the full industrial or biological strength necessary to accomplish useful work in the world. Indeed a well-worn argument which this paper too supports is the view that classical knowledge-based systems will not scale up. Are there any reasons to suppose that the behaviour-based or enactive views espoused here will scale up?

In developing suitable terminology for discussing the varieties of knowledge which might be employed by people, bats, beetles, robots, and thermostats, we have been led to a view of mind and knowledge which extends beyond the brain into the physiology/technology of the creature's body, and beyond that, into the world, and into other creatures. While the brain may be the psychopoetic organ, it is not the entire host of mind. Many people from a variety of other viewpoints have been led to a similar view.

- The cyberneticist Bateson argued that since a creature interacts with its environment via lots of servo loops of various levels, in which information is constantly circling around the world → sensor → brain → effector → world loop, mentality is immanent in the whole inseparable dynamic ensemble of creature/local-world [Bateson 71].
- The modern cyberneticists Varela and Maturana have developed this view into their theory of autopoesis [Maturana & Varela 80].
- Before modern computer science and control theory had developed, the theoretical biologist von Uexküll was arguing for a similar conclusion based on his biological theory of meaning [von Uexküll 40].
- The philosopher Putnam has used his parallel worlds arguments to deduce that mind is not entirely in the brain, or indeed, in the creature [Putnam 75,Putnam 87].
- The modern teleosemanticist Millikan is developing a philosophically and scientifically principled way of grounding meaning in objectively discernible biological function [Millikan 84].
- Phil Agre [Agre 97], Andy Clark [Clark 98], and Ron McClamrock [McClamrock 95] argue from detailed considerations of philosophy of mind, psychology, and computational theory, that a shift is required away from considering the innards of mind towards focusing on patterns of interaction with the world.

I have so far left unanswered the interesting question of whether the **understandingB** I have been concerned with here is a step towards **understandingC**. That is not a question which can be easily answered by studying the entrails of robots. I note, however, that there is a recent vigorously developing programme of research into the natural evolution of human **knowledgeC** which *does* base itself on previously evolved animal **knowledgeB** of the specifically interactive embodied and embedded kind characteristic of behaviour-based robotics which I have been considering here.

- Susan Hurley [Hurley 98] ranges carefully over modern psychology and neurophysiology to argue that when **understandingB** is properly rooted in Gibsonian "affordances" [Gibson 79] and the kind of "embeddness" and "embodiment" characteristic of behaviour-based robotics then it is indeed a step towards **understandingC**, and finds philosophical roots of these ideas as far back as Kant, Aquinas, and further.
- Nunez & Freeman provide in an edited collection of papers a sample of the range of views and research contributing to this modern scientific development of phenomenology and cybernetics [Nunez & Freeman 99].
- Having mentioned cybernetics I should include the modern inheritors of the mantle of General Systems Theory, such as the complex systems research programmes at the Complexity Research Group of Virginia Commonwealth University, the Santa Fe Institute, and many more, which are elucidating general principles and architectures with the assistance of the new discipline of experimental mathematics facilitated by the modern computer.

This is far from an exhaustive list. There is clearly a multi-disciplinary convergence approaching the problem of naturalising mind (explaining how mind can arise by natural processes from matter) from many directions, and a growing recognition that mind is a dynamic multi-layered manifestation which extends beyond the brain and beyond the creature into the creature's local interactive world.

All this provides some hope that the kind of work currently being undertaken in the behaviour-based robotics tradition has a good future ahead of it, rather than being confined to a dead end niche in the evolutionary history of artificial mind as insects have been in the biological evolution of mind, and as Winograd & Flores argue classic knowledge-based AI systems are [Winograd & Flores 86].

11 CONCLUSION

Now that we are seriously working towards the business of implementing mind in machinery, it is clearly necessary to be able to contrast and compare the cognitive capacities of nonconscious creatures, both natural and artificial, and of cognitive subsystems. This requires a vocabulary with which we can clearly distinguish the kinds of knowledge possessed by calculators, robots, birds, and people. This paper has developed a simple example of such a vocabulary, and shows how it can be extended.

- the **A** or **A**lan Turing version which simply has to display the right behaviour;
- the insistence that the behaviour be produced by the right kind of nonconscious mechanism, the **B** or **B**rian Smith version;
- and finally the **C** or **C**onscious version, which insists on the involvement of consciousness.

This kind of tripartite distinction can also be made of meaning, understanding, decision, and a variety of other mentalistic terms. This tripartite division of knowledge was based on considering criteria for accepting the *presence* of knowledge in a creature. When we turned to look at knowledge from the other side, concerning how to *implement* it in a creature, we found a different fourfold classification of **symbolic/explicit**, **symbolic/implicit**, **tacit/procedural**, and **tacit/dispersed**. Considering the **dispersed** category of **knowledgeA** led us to relate these two classifications of knowledge via the technology domain ladder of creature construction, which in turn suggested further interesting distinctions and generalisations.

Debates about the nature of thinking and knowledge, to what extent machines might "think" or "know", have often been concerned with settling on one particular definition of "knowledge" or "understanding" and criticising others. The purpose of this paper is to argue the utility of developing a range of definitions, a variety of types, of knowledge, understanding, etc., and exploring the properties and inter-relations of the whole range.

The development of these concepts of understanding, knowledge, etc., naturally led to a view of mind which extended beyond the brain (or computer) into the interactive complex of creature and local world. That a number of biologists, psychologists, and philosophers of mind are also finding that a congenial view suggests that, unlike the classical knowledge-based AI research programme, based on a computational model of mind, this one may have the potential to scale up.

References

[Agre 97] Phil Agre, *Computation and Human Experience*, Cambridge University Press 1997.

[Bateson 71] Gregory Bateson, *The Cybernetics of "Self" : A Theory of Alcoholism*, in **Psychiatry**, Vol 34, no 1, pp 1-18, 1971; reprinted in *Steps to an Ecology of Mind*, Ballantine Books, NY, 1972.

[Brooks 86] Brooks, R.A, *Achieving Artificial Intelligence through Building Robots*, AI Memo 899, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, May 1986.

[Chalmers 96] Chalmers, David John, *The conscious mind : in search of a fundamental theory*, New York Oxford : Oxford University Press, 1996.

[Clark 98] Andy Clark, *Being There : Putting Brain, Body, and World Together Again*, Bradford Books, MIT Press, 1998.

[Clark 99] Andy Clark, *Visual Awareness and Visuomotor Action*, in "Journal of Consciousness Studies", **6**, No. 11–12, 1999, pp. 1-18.

[Gibson 79] J J Gibson, *The Ecological Approach to Visual Perception*, Boston, Houghton Mifflin, 1979.

[Harnad 87] ed Harnad, Stevan, *Categorical perception : the groundwork of cognition*, Cambridge : Cambridge University Press, 1987.

[Haugeland 98] John Haugeland, in *Having Thought*, "The Intentionality All Stars", Harvard, 1998.

[Haugeland 85] John Haugeland, *Artificial Intelligence: The Very Idea*, MIT Press, Cambridge, MA, 1985, a Bradford Book.

[Hofstadter & Dennett 82] *The Mind's I*, eds D.R. Hofstadter and D.C.Dennett, Penguin, 1982, ISBN 0-14-00.6253-X, (paperback).

[Humphrey 92] Nicholas Humphrey, *A history of the mind*, London, Chatto and Windus, 1992.

[Hurley 98] Susan Hurley, *Consciousness in Action*, Harvard University Press, 1998.

[Malcolm et al 89] Malcolm, C, Smithers, T, and Hallam, J, *An Emerging Paradigm in Robot Architecture*, invited paper at the Intelligent Autonomous Systems Conference (2) in Amsterdam, Dec 11-14, 1989; also available as Edinburgh University DAI RP 447.

[McClamrock 95] , Ron McClamrock, *Existential Cognition : Computational Minds in the World*, University of Chicago Press 1995.

[Maturana & Varela 80] Maturana, Humberto R., Varela, Francisco J., *Autopoiesis and cognition : the realization of the living*, Dordrecht London : Reidel, 1980, Boston studies in the philosophy of science ; v. 42

[Michie 73] , *Britain's Bicycle Shed*, New Scientist Feb 22nd 1973.

[Millikan 84] Millikan, Ruth Garrett, *Language, thought and other biological categories : new foundations for realism*, Cambridge, Mass. London : MIT, [1984], A Bradford book.

[Nagel 74] Nagel, T, *What is it like to be a bat?* Philosophical Review 4:435-50. Reprinted in Mortal Questions (Cambridge University Press, 1979), 1974.

[Nevins and Whitney 78] Nevins, J.L., and Whitney, D.E., *Computer Controlled Assembly*, Scientific American, February, 1978.

[Newell & Simon 76] A Newell and H Simon, *Computer Science as Empirical Enquiry*, in *Mind Design*, Haugeland (ed), Bradford Books, MIT Press 1981.

[Newell 81] Allen Newell, *The knowledge level*, Department of Computer Science, Carnegie-Mellon University, 1981 SERIES :CMU-CS- ; 81-131

[Nunez & Freeman 99] Rafael Nunez & Walter J. Freeman, eds, *Reclaiming Cognition : the primacy of action intention and emotion*, Imprint Academic, 1999.

[Russell and Norvig 95] Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

[Pfeifer 96] Rolf Pfeifer, 'Building "Fungus Eaters": Design principles of autonomous agents', in: "FROM ANIMALS TO ANIMATS 4", Fourth International Conference on Simulation of Adaptive Behavior, Cambridge, MA: The MIT Press/Bradford Books (1996).

[Putnam 75] Putnam, H. 1975. *The meaning of 'meaning'*, Minnesota Studies in the Philosophy of Science 7:131-193. Reprinted in Mind, Language, and Reality (Cambridge University Press, 1975).

[Putnam 87] Putnam, H. 1987. Meaning, other people, and the world. In Representation and Reality. MIT Press.

[Pylyshyn 86] Zenon Pylshyn, *Computation and Cognition : Toward a Foundation for Cognitive Science*, MIT Press 1986.

[Schank and Abelson 77] Roger C. Schank & Robert P. Abelson, *Scripts, plans, goals and understanding : an inquiry into human knowledge structures*, Hillsdale, N.J. : L. Erlbaum Associates, 1977.

[Searle 80] John Searle, *Minds, Brains, and Programs*, Behavioral and Brain Sciences, 3, p417-457, 1980.

[Shannon and Weaver 49] Claude E. Shannon & Warren Weaver, *The mathematical theory of communication* University of Illinois Press, 1949.

[Smith 82] Smith, B.C., *Reflection and Semantics in a Procedural Language*, PhD dissertation, Report MIT/LCS/TR-272, MIT, Cambridge, MA, 1982. See also, chapter 3 of Brachman, R.J. and Levesque, R.J., Eds, **Readings in Knowledge Representation**, Morgan Kaufmann, California, 1985, pp31-40.

[Turing 50] Turing, Alan, *Computing Machinery and Intelligence*, Mind 59, October 1950, pp433–60, reprinted in *Computers and Thought*, eds Feigenbaum and Feldman, McGraw-Hill 1963.

[von Uexküll 40] Uexküll, J. von (1864/1944), *The Theory of Meaning*, English translation, Semiotica, 42-1, (1982) pp25-82.

[Winograd & Flores 86] Terry Winograd & Fernando Flores, *Understanding computers and cognition : a new foundation for design*, Reading, Mass. ; Wokingham : Addison-Wesley, 1986.

[Varela 91] Varela, F.J., Thompson, E., and Rosch, E., *The Embodied Mind*, MIT Press, 1991.

[Zohar 90] Zohar, Danah, *The quantum self : a revolutionary view of human nature and consciousness rooted in the new physics*, London : Bloomsbury, 1990.