Grand Challenge 5:[1]

The Architecture of Brain and Mind

Integrating Low-Level Neuronal Brain Processes with High-Level Cognitive Behaviours, in a Functioning Robot[2]

Moderator: Aaron Sloman[3]
(A.Sloman@cs.bham.ac.uk)

[1]This is one of a number of proposals being discussed in the framework of the UK 'Grand Challenges' initiative of the UK computing research committee (UKCRC).
For more information see **http://www.cs.bham.ac.uk/research/cogaff/gc/**

[2]This document expands on the draft produced by the original moderator for GC5, Mike Denham (University of Plymouth), in January 2003. Since then discussions that have occurred since then in the GC5 mailing list and elsewhere, especially the GC5 workshop at De Montfort University on 5th January 2004, and the UKCRC Grand Challenges Conference at Newcastle, 29-31 March 2004. Thanks to Mark Lee (Aberystwyth), for some of the wording of the introduction.

[3]School of Computer Science, The University of Birmingham. http://www.cs.bham.ac.uk/~axs/

Contents

1 Introduction **3**

1.1 The time is ripe. . . . 3
1.2 Older approaches . . . 4
1.3 The need for mechanisms . . . 4
1.4 A more recent alternative . . . 5
1.5 Knowledge about artificial virtual machines and their implementation . . . 6
1.6 Structural relations between virtual and physical machines . . . 6
1.7 Benefits of a hybrid methodology . . . 7
1.8 Layers of implementation . . . 8
1.9 Knowledge about natural virtual machines and their implementation . . . 8
1.10 Limitations of learning about correlations between levels . . . 8
1.11 Studies of high level human virtual machines . . . 9
1.12 Keeping an open mind . . . 10

2 The Challenge **10**

2.1 The key role of computing . . . 11
2.2 Scientific objectives . . . 11

3 Required scientific advances in computing **12**

3.1 Practical applications . . . 12

4 What kind of grand challenge is it? **13**

4.1 Milestones and backward-chaining . . . 14

5 How will it be done? **15**

5.1 Alternative targets . . . 15
5.2 International collaboration . . . 17

1 Introduction

> *What is the most powerful and most complicated computer on the planet? Wrong! it's not a machine you can buy for millions of dollars, it's the amazing system that we all own, the few kilos of grey and white mush in our heads.....*

Biological information processing systems produced by evolution still far outstrip both our understanding and our practical achievements: there are deep gaps in our theories and in our engineering capabilities.

In the hope of reducing both gaps we shall look closely at two of the most impressive products of evolution: human *brains* and human *minds* – and attempt to construct a combined vision of how they work demonstrated in a robot that goes far beyond what current systems can do:

- *Brains*, the contents of our skulls, are composed of extraordinarily intricate, self-organising, physical structures, performing many tasks in parallel at many scales, from individual molecules to large collections of cooperating neurones or chemical transport systems.
- *Minds* are more abstract and contain ideas, perceptions, thoughts, feelings, memories, mathematical knowledge, motives, moods, emotions, reasoning processes, decisions, motor control skills and other things that cannot be seen by opening up skulls. Yet their existence and their power to do things depend on all the 'wetware' components that make up brains.

The end goal: a fully functional robot implemented using an artificial brain built out of components simulating low level functions of animal brains will not be achieved in the foreseeable future. But

- a robot using more abstract models of higher level brain functions and combining many kinds of functionality,

  along with
- parallel demonstrations of the plausibility of the claim that those brain functions could be implemented in mechanisms simulating very low level brain mechanisms,

could be achieved in 15 to 20 years.

1.1 The time is ripe.

The last twenty years have seen an explosion in the application of molecular biology, genetics, and cell biological techniques to the problems of neurobiology, and a growth in neurobiological experimental research which has dramatically increased our understanding of the nervous mechanisms and their functions. However, insofar as the primary function of the nervous system is to gather, represent, interpret, use, store and transmit information (including information about the environment and information about internal states and processes, and including both factual and control information), neuroscience is inherently a computational discipline, in the broad sense of 'computational' that covers all information processing.

So, despite the insights neurobiology provides, a mature science of the brain ultimately also requires a deep understanding of how information is represented, organised, manipulated and used within the structures of the nervous system, and how such brain processes create the high-level cognitive capabilities which are manifest in the human mind.

In addition, in a world that day-by-day becomes increasingly dependent on technology to maintain its functional stability, there is a need for machines to incorporate correspondingly higher and higher levels of cognitive ability in their interactions with humans and the world. Understanding the principles of brain organisation and function which subserve human cognitive abilities, and expressing this in the form of an information-processing architecture of the brain and mind, will provide the foundations for a radical new generation of machines which act more and more like humans. Such machines would become potentially much simpler to interact with and to use, more powerful and less error-prone, making them more valuable life-companions, whether for learning, information finding, physical support or entertainment. They might even be able to recognize even the best disguised spam email messages as easily as humans do!

Despite the importance of these potential practical applications, the primary focus of this grand challenge project is on developing scientific understanding including a new unification of concepts, theories and techniques that will answer age-old questions about the nature of mind and the relation between mind and body. It will also address the question whether currently known forms of computing machinery provide an adequate basis for replication of all the major functions of brains, or at least their cognitive functions, or whether new kinds of computers are required, as some have claimed.

This is the sort of scientific curiosity that makes us want to know what exists in the far reaches of the universe, or how space and the objects in it were created billions of years ago, or how the huge multitude of organisms on earth evolved. If we thereby gain new technical skills or practical benefits as a result of such advances in understanding, that is a very welcome bonus. In this case, many bonuses are inevitable.

1.2 Older approaches

This is not the place for a comprehensive history of theories of the nature of mind and the relation between mind and body. There are standard philosophical introductions to the topic in, for example, (Campbell 1970; Chalmers 1996), among hundreds of others. However, it is worth mentioning that until the early twentieth century most people attempting to understand how brains and minds are related were forced to choose between two equally unacceptable options: *the naive materialist view* that minds do not exist, and only brains and bodies do, and *the naive dualist view* that minds exist independently of brains and can survive the destruction of the body because they are composed of a separate type of 'substance'. There were more subtle theories, such as that mind and matter were both merely aspects of a 'neutral' kind of stuff that was neither mind nor matter, and the twentieth century 'behaviourist' vision of minds as mere sets of relationships between inputs and outputs of brains (or combined brain-body systems), so that the study of mind was reduced to the study of contingencies of behaviours of various kinds. Variations of this theme are found in a range of behaviourist theories, and simplistic 'functionalist' theories in the cognitive sciences.

1.3 The need for mechanisms

All those theories lacked explanatory power, insofar as they failed to explain how anything *worked*. For instance they could not explain how light waves impinging on our eyes could produce perception of chairs around a table (including partially occluded chairs) or water flowing in a stream, how desires arise, how inferences are made, how plans are formed, how decisions are taken, how moods and emotions come and go, how a child learns to talk, and to think about

numbers, and so on.

In order to fill this explanatory gap, many attempts were made to postulate *mechanisms* producing mental processes. Sometimes it was thought that if certain physical mechanisms produced mental states and processes in humans, then by replicating those mechanisms in artifacts, it would be possible to make machines that could think, perceive, have desires, and so on, without understanding how the mechanisms actually worked, as in the case of ancient stories about 'golems'. But we already know how to make new thinking machines that we don't understand — we do it all the time.

The invention, or discovery, of new types of machines led to new models of what minds were and how they worked. For instance, in the early days of electronics, analogies were drawn between brains and telephone exchanges, leading to a view of mind as some sort of pattern of communication activity between portions of the body, even though nobody had ever encountered a telephone exchange that could learn to talk, fall in love, or make plans.

Another kind of explanatory notion came from control theory (Wiener 1961) insofar as homeostatic mechanisms involving feedback loops could be seen to be *goal-directed*, and therefore partly like humans and other animals. The electro-mechanical 'tortoise' of W.Grey Walter [4] was a product of such ideas. Increasingly elaborate versions of such theories were developed, though sceptics could not believe that any number of control circuits with changing numerical values for pressures, currents, voltages, forces, velocities and other physical quantities could explain the occurrence of structured thoughts and inferences, for instance working out that if a polygon has three sides its internal angles must add up to a straight line, or concluding that a forecast of rain makes it prudent to go out with an umbrella.

1.4 A more recent alternative

A new, deeper, more general form of explanation became available as a result of developments in the last half century in computer science and software engineering, leading to the design and implementation of a wide range of working hardware and software systems including operating systems, office automation systems, flight control and plant-control systems, email systems, database systems, online reservation systems, and many more.

In particular, research in AI (including computational cognitive science) increasingly supplemented the arithmetical capabilities of computers used in simpler control systems with symbolic capabilities of many sorts, using algorithms and forms of representation that enabled computers to perform many tasks previously done only by humans, including playing board games, finding mathematical proofs, answering questions, checking for hazards, and controlling machines or even factories.

All such systems include both *physical machinery* (which occupies space, consumes energy, can be weighed, moved, inspected with microscopes etc.) and also working *virtual machines* containing data-structures, algorithms, scheduling mechanisms, interrupt handlers, file systems, priorities, permissions, etc., that are *implemented* (or *realised*) in the physical machines. Both sorts of machines exist, produce effects, have internal structures and processes, can go wrong, and can be fixed (sometimes). The physical and virtual machines have to be understood in their own terms by anyone trying to understand complex information-processing systems.[5]

[4] http://www.epub.org.br/cm/n09/historia/documentos_i.htm

[5] It is important that we are here referring to *running* virtual machines that are actually processing information,

1.5 Knowledge about artificial virtual machines and their implementation

People who design, develop, debug, and maintain physical and virtual machines use different ontologies,[6] different theories and different sets of creative skills. Typically a software engineer cannot fix hardware bugs and an electronic engineer cannot fix software bugs on the same running computing system. There is, however, a subset who can think at both levels and whose theoretical and practical work, e.g. in the design of compilers, new computer architectures, and low level operating system or networking mechanisms, ensures that the interfaces between virtual and physical machines work as required, e.g. so that when other designers using the system do not make mistakes, execution of machine-code instructions for manipulating sets of switches treated as bit patterns produces the right virtual machine events, e.g. finding spelling mistakes in a document, displaying a railway time-table correctly on a screen, finding errors in a mathematical proof, or selecting the correct control signals for rocket machinery during takeoff.

1.6 Structural relations between virtual and physical machines

The existence of humans and other animals with capabilities still unmatched by the machines we know how to design, shows that evolution 'discovered' many designs and mechanisms about which we are still ignorant. Some believe that the best or only way to understand those products of evolution is to investigate their internal mechanisms and their behaviours in great detail.

This grand challenge proposal suggests that that 'traditional' scientific approach may leave us ignorant of the best ontology to use in studying natural systems, and offers a broader research strategy: we can make more progress by combining the study of biological systems with what we have learnt from our own design efforts, just as advances in mathematics that do not arise from observations of physical phenomena may turn out later to be useful in physics.

One of the most important facts to have emerged in the last half century, which is not widely known, is that there need not be simple relationships between what exists or can happen in a *virtual* machine and the physical structures, processes and states in the *physical* machines in which they are implemented. The structure, complexity and variety of components of a virtual machine can change while the number and variety of components of the underlying physical machine remain fixed. For instance, by installing different software systems on the same machine we can change the operating system on a PC from Windows to Unix or Linux thereby changing a machine that supports only one user at a time to one on which different users can be logged in simultaneously running different virtual machines, even though no machine components have been changed. In fact both operating systems can be installed permanently in the machine, so that a simple selection by the user when the machine restarts determines which operating system runs. There are even

taking decisions, etc. and not to the kind of *mathematical abstraction* that is often called a virtual machine, such as the Java or Prolog virtual machine where these are static structures, which merely describe or specify classes of running virtual machines. These mathematical entities can no more cause events to happen than the number 3 can. In contrast, the running instances frequently make things happen, control how things happen, or prevent things from happening.

[6] In philosophy, the word 'ontology' refers to at least two different sorts of things: (a) the most general study of what exists and (b) a set of general assumptions about what kinds of things exist presupposed by some community. In AI and software engineering the word now often refers to a specification (formal or informal) of some aspect of reality presupposed by a system which makes use of information about the environment (including other agents and itself). E.g. an ontology might specify the kinds of objects, properties, relationships, events, processes and causal relationships that a machine will perceive, reason about, communicate about, act on, etc. In that sense, the designer of the machine also uses an ontology, normally one that includes the ontology required by the machine along with the kinds of things a designer needs to know about such machines.

software systems that allow both operating systems to run at the same time, without changing the physical construction of the machine.

Of course, switching from one virtual machine to another involves different detailed physical processes when the machine runs, even if there is no re-wiring or replacement of physical components: the changes involve different sequences of state-changes in large numbers of tiny switches in the computer.

It is also possible for a virtual machine to continue running while some of the underlying physical machine is replaced. E.g. in computers with 'hot-swappable' components, it may be possible to replace memory module or a power-supply or hard drive, with new ones that have a different physical design, while the virtual machine goes on running as before.

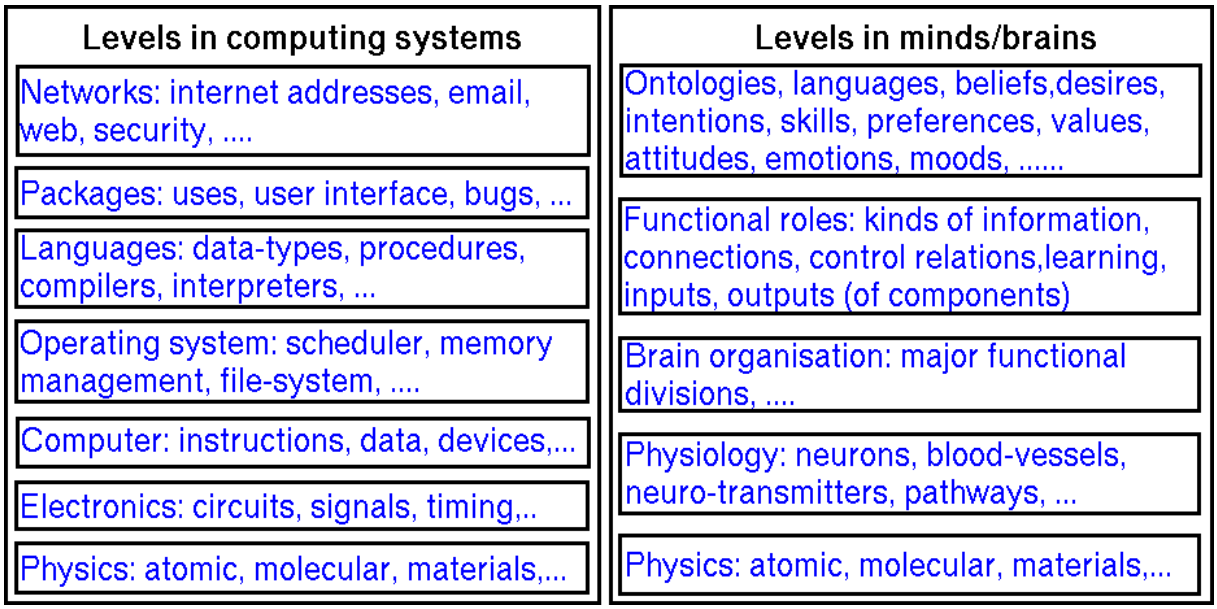

Figure 1: *Levels of implementation — virtual machines at the top and at intermediate layers, and physical machines at the bottom. These are not meant to be accurate summaries, merely indicators of the concept of layered virtual machines in man-made and natural systems. There is no simple correspondence between levels on the left and on the right: the juxtaposition merely indicates that both involve several levels of implementation: those on the left produced by human engineers, those on the right by evolution, development, learning, and cultural evolution.*

1.7 Benefits of a hybrid methodology

If we adopt the proposed hybrid approach combining what we have learnt from work on computers with results of biological research, this may enable us to ask questions that might not occur to researchers of one sort or the other. E.g. we can ask whether evolution 'discovered' the usefulness of being able to switch the virtual machine running on a physical machine (e.g. having a re-usable part of the brain that can be used for multiple different purposes) or being able to switch which physical machine a virtual machine runs on (e.g. perhaps running processes on one physical machine while a skill is being developed and later transferring it to another part of the brain as expertise is achieved).

There are also many other advantages that come from the fact that partial results from each explanatory level can usefully constrain the search for good explanations at other levels.

1.8 Layers of implementation

As indicated in Figure 1, we have learnt that within both the virtual realm and the physical realm there can be several layers of implementation, for instance when a plant-control virtual machine is implemented using the virtual machine for an operating system which uses the run-time virtual machine of a programming language, which uses the bit-manipulating virtual machine of something like a sparc or pentium processor. Likewise the physical digital circuitry may be implemented in physical and chemical mechanisms which are describable both at molecular levels and at sub-atomic quantum mechanical levels, and perhaps further levels, depending on how physics develops in the future.

All of this understanding of how layers of implementation in virtual and physical machinery work has come from years of design and implementation of ever more sophisticated working systems: we understand them at least well enough to build them, debug them, and improve or extend their functionality. It might have taken us much longer to develop the appropriate concepts and explanatory theories if we had merely *observed* and *experimented on* such machines imported by an alien culture, without ourselves being involved in their design and development. An aspect of this grand challenge project is discovering whether what we have learnt from designing many sorts of layered virtual machines may help us ask the right questions when investigating machines produced by evolution.

Another aspect is discovering whether the versions produced by evolution can teach us about kinds of virtual machine that we have not yet invented.

1.9 Knowledge about natural virtual machines and their implementation

At present there is relatively little understanding of how most mental entities, events, and processes (e.g. percepts, decisions, plans, beliefs, desires, etc.) are related to the underlying physiological, physical and chemical mechanisms and events and processes therein.

This is not for want of significant recent advances in brain science. Up until very recently, our knowledge of the mechanisms of the brain has been very sparse and limited in depth. It was only in 1952, that A.L. Hodgkin and A.F. Huxley first described the voltage clamp method for measuring neuronal response which has formed the basis for much of the neurobiological experimental investigations since that time. However, in the last ten or so years methods of imaging of living human brains (PET, fMRI) have provided a wealth of new knowledge, about which parts of brains are involved in various kinds of cognitive functions. One of the most recent techniques to be developed is an MRI-detectable, neuronal tract-tracing method in living animals, which recently demonstrated MRI visualisation of transport across at least one synapse in the primate brain. Further research will include studies of development and plasticity and is likely to provide new valuable information about brain anatomy.

1.10 Limitations of learning about correlations between levels

Brain researchers are constantly developing new means of investigating physical processes in brains associated with various kinds of mental processes and external behaviours, so that there is growing accumulation of evidence regarding correlations between brain occurrences and mental

occurrences, e.g. between increased blood flow or electrical activity in a particular part of the brain, and the existence of various mental states and process, such as seeing something, taking a decision, or performing an action.

We can expect continuing development in all aspects of neuroscience, using technical innovations in neurobiological observation and measurement (partly based on the availability of vastly greater computer power) and also techniques at genetic, molecular and cellular levels. The spatial and temporal resolution of brain-scanning devices will continue to increase, allowing ever more research into which brain locations are involved in particular mental functions.

But such correlations do not help us understand what the mental states and processes are, at a functional level (for instance what is involved in understanding a joke, or seeing how a machine works), or how they are implemented in brains: knowing *where* something happens does not tell us *how* it happens. Moreover it leaves open the possibility that additional processes elsewhere, not detected by the scanning techniques available, are just as important as those detected, in the way that the significance of changes in bit patterns in a computer's memory or in its CPU depends on many other states and processes in other parts of the machine. The very same sequence of bit-level operations in a CPU might in one context be part of an arithmetic process of adding two numbers, and in another context part of a process of relating information in two memory locations.

Nevertheless there has been and will continue to be a steady growth in information about detailed brain mechanisms and what happens when damage or disease interferes with normal functioning. This provides new opportunities for increasing our understanding of the diversity of 'low level' mechanisms in the brain, and how they interact as components of a complex integrated system, thereby bringing us closer to understanding how the physiological mechanisms are capable of supporting higher level cognitive and other capabilities.

But we cannot simply depend on such 'bottom-up' research to answer all the important questions. One reason for the difficulty of understanding the relationship between physical and virtual machines is that the higher level functions, such as perception, learning, problem-solving, communication, control of actions, etc. depend on the emergent functional properties of vast systems of neural circuits and their interactions with the environment, where the emergent behaviours are not simple mathematical combinations of the behaviours of the individual components, any more than the behaviour of an operating system, or spelling checker in a computer is a simple mathematical function of the behaviour of the individual transistors, connectors, etc.

1.11 Studies of high level human virtual machines

There is of course, a vast amount of empirical research on structures, processes, and causal interactions in higher level virtual machines in humans and other animals, for instance research attempting to investigate stages at which children develop and use concepts relating to mental states and processes, such as *belief*, *desire*, *happy*, *angry*, etc., and decades of research on reasoning, learning, communication, motivation, emotions, and other 'virtual machine' phenomena. However, in general such research is done by people who do not put forward theories about how the machines work: they are mainly concerned about the conditions under which various things happen, and how speed, and other performance features such as error rates, types of errors, etc., vary with circumstances.

Even the recent attempts to link virtual machine states and processes with brain states using new brain imaging technology, mostly establish correlations, without explaining how anything works

(like finding which bits of a computer's hardware are active when a spelling checker runs, or when email is being sent).

There is a small subset who attempt to design and build working computer models to check whether their theories really do have the claimed explanatory power. That design and implementation process in itself often leads to the discovery of new problems and new theoretical options which advance our understanding. But such models typically assume the availability of a certain level of virtual machinery on top of which the proposed mechanisms can be built, and does not explain how such virtual machinery might be implemented in brains. For example, even most computational neural nets have very little in common with biological neural mechanisms, insofar as they fail to capture important details of the ways in which individual neurons work as complex information-processing systems extended in space. Some researchers, though not all, believe that such artificial neural nets nevertheless summarise an important level of virtual machinery implemented in the neural mechanisms.

1.12 Keeping an open mind

We should keep an open mind as to whether the mechanisms used in current working computational models might turn out not to be implementable in brains, in which case the proposed theories, despite their explanatory and predictive power cannot be right *at that level*. It is possible that, by finding out more about previously un-thought of kinds of virtual machines that *can* be implemented in brains, we shall be led to discover more varied frameworks for proposing higher-level explanatory mechanisms, helping us on the probably unending journey towards true theories.

This does not imply that work based on existing, well understood, computational virtual machines, such as artificial neural nets, or problem solving systems like SOAR (Laird *et al.* 1987) or ACT-R[7] is completely wasted: for even if they are not correct models of intermediate virtual machines in human minds, by working with them we can learn more about *requirements* for adequate theories by exploring designs to find out what sorts of things proposed models can and cannot do. Insofar as there are some subsets of cognitive functioning that those models do seem to be able to explain well, e.g. details of processing in natural language, or certain kinds of problem solving and the classes of errors that arise, these theories will help us to specify some of the *requirements* for more biologically accurate virtual machines that support the full range of phenomena to be explained.

In contrast, merely working upwards from the study of the lowest-level components may not suffice to enable us to understand how the macroscopic behaviours supporting higher-level virtual machines arise. For instance it may not lead us to the best ontology for describing the macroscopic behaviours.

Our claim is that progress may be considerably accelerated if we have an independent route to characterising the nature of the 'systems-level' behaviours: so we should work bottom-up, top-down and middle-out in parallel.

2 The Challenge

Our grand challenge then is to develop concepts, theories and tools that can be tested by producing a fully functional robot combining many kinds of human competence, perceiving, acting, having motivations and emotions, learning, communicating, and to some extent understanding itself,

[7] http://act-r.psy.cmu.edu/

in parallel with attempting to demonstrate through further working models that the kinds of mechanisms used could be implemented in biological brain mechanisms.

This will require us to combine advances in understanding of systems designed by us:

- understanding of varieties of information processing in high-level virtual machines, investigated by researchers in computer science, AI, computational intelligence, cognitive science and software engineering,
- understanding of physical machines for implementing computational systems

with advances in our understanding of natural systems,

- in neuroscience and the relevant physical/chemical sciences
- in ethology, psychology, linguistics and other sciences which study what humans and other animals can do.

Although much work can be done separately in the two streams, combining them in new ways is essential for major advances in both. This is also the assumption of the UK Foresight Cognitive Systems project,[8] though our proposal goes further in insisting on the combination of the many components of cognitive systems within a functioning robot, on a longer time scale.

2.1 The key role of computing

The really long term *computing* grand challenge will be to develop a succession of increasingly comprehensive working models of both mental functioning and brain mechanisms, culminating in a system where the working models of low-level neural mechanisms are demonstrated to be capable of supporting a variety of mental functions, including perception, learning, reasoning, planning, problem-solving, motivation, emotions, action control, self-awareness and social interaction, including communication in a human language. These diverse functions should be fully integrated in a single system, operating concurrently and interacting when appropriate. Insofar as the system is a model of a human or animal with physical sensors acting in a physical environment, the model should be some sort of robot.

It is unlikely that that level of integration can be achieved in 15-20 years. But it should be possible to demonstrate a working system in which the virtual machines at least match the biological versions at certain level of abstraction. At the same time the project will have generated new requirements for explanatory mechanisms in brain science and many of them should have been modelled within two decades.

2.2 Scientific objectives

This is not primarily an engineering project to design a useful robot, but a scientific endeavour requiring not only that the working models should reflect what is learnt from the other disciplines but also that the process of developing them should inspire new theoretical advances and empirical research in those disciplines.

So one indicator of success will be increasingly deep and fruitful collaboration between those working on this challenge to build working systems, and researchers in the other relevant disciplines, including biology, neuroscience, psychology, linguistics, social science and philosophy, with a two-way flow of knowledge, much as developments in mathematics and development in physics went hand-in-hand as both disciplines evolved over several centuries.

[8] http://www.foresight.gov.uk/

3 Required scientific advances in computing

This will not be a simple matter of using existing know-how in computer science, software and hardware engineering, and AI to implement independently formulated theories. It will require major advances in hardware and software design techniques, including new formalisms for specifying requirements and designs at various implementation levels. This will include designs for complex systems combining many kinds of functionality, operating in parallel with and in close interaction with both other internal sub-systems and animate and inanimate entities in a complex physical and social environment.

It is very likely to require the development of:

- *New forms of representation* for use in various sub-systems of the working models (e.g. visual perception, auditory perception, tactile perception, language understanding, long term memory, action control, proprioceptive feedback, motivational states, conflict resolution, mood control, self-understanding, and understanding of other information processors)
- *New ways of specifying designs* for such complex systems, including novel architectures specified not only in terms of the low-level neuronal mechanisms used, but also in terms of the various combinations of high-level functionality, including reactive capabilities, deliberative capabilities, self-reflective capabilities, motivational control systems, etc.
- *New techniques for getting from design specifications to working systems*, e.g. new kinds of system-compilers, new kinds of configuration-management systems, new kinds of tests for mismatches between designs and working systems, etc.
- *New ways of specifying requirements* against which designs and implementations are to be tested (e.g. requirements for perception of affordances[a], requirements for social competence, requirements for self-understanding, requirements for extended periods of human-like learning and development, requirements for creativity, requirements for appropriate forms of physical behaviour, etc.)
- *New techniques for checking relationships* between requirements and designs, and between designs and implementations. In conventional computing there is generally a clear separation between the work of hardware designers who produce components used in computers, including memories, CPUs, buses, device controllers, network controllers, etc., and the work of software designers who generally assume that the hardware will perform according to specification and do not need to know *how* it meets the separation. It is not clear that a similar clear separation will be possible in connection with a system in which high-level cognitive functions are implemented in low-level brain-like mechanisms. For instance, for a robot moving about in a natural environment there may be multiple timing constraints at all levels, or high-level action-control functions using feedback involving low-level sensory transducers.

[a] Perceiving affordances includes not merely perceiving objects properties and relationships in a scene, but also perceiving opportunities for action (positive affordances) and obstacles to action (negative affordances). The idea comes from the work of (Gibson 1986).

3.1 Practical applications

The general techniques and specific designs developed for this purpose should provide major new opportunities for practical applications of many kinds, including intelligent robots and virtual agents which incorporate a significant subset of human cognitive capabilities. These might include

new kinds of intelligent self-monitoring operating systems and computer networks, intelligent personal assistants, intelligent disaster support systems, and intelligent aids to scientific research in many disciplines where complex experiments have to be controlled and huge amounts of data interpreted, for instance.

Advances on the grand challenge project will also have many practical applications apart from building new useful systems. Insofar as it enhances our understanding of natural intelligent systems, it could provide deep new insights into practical problems concerned with humans, for instance trying to understand or alleviate a range of human mental disorders; and helping educators trying to understand what goes on when children learn or fail to learn or develop as expected.

For instance, consider the task of designing and testing working robots able to develop their understanding for numbers from the level of a child starting to learn to count to the understanding of a bright ten year old who understands that there are infinitely many integers, that there are positive and negative integers, fractions, etc., and that numbers can be applied in many practical tasks involving both discrete sets of objects (e.g. people and chairs) and continuously varying quantities, such as lengths, areas, volumes, weights, etc. This is likely to reveal that the current purely *mathematical* understanding that we have about these things leaves huge gaps in our understanding of the processes of cognitive development required for such understanding. If we can begin to fill those gaps, the results may include far-reaching implications for successful teaching of mathematics in primary schools.

Likewise if we can accurately model varieties of affective states and processes (motivation, preferences, conflict resolution, emotions, attitudes, moods, values, including self-categorisations in pride, guilt, shame, etc.), and if our models can be used to demonstrate processes by which various sorts of disorders can arise, some produced by low-level damage in implementation machinery, some by developmental errors during growth of architectures, some requiring specific gaps or errors or imbalances in high-level virtual machine operations, then we may acquire deep new insights into good ways of categorising, preventing and treating a wide variety of disorders, providing a rich armoury of new conceptual tools and modelling tools for practitioners in various kinds of therapy and counselling.

However the potential applications are not the *main* focus of this challenge: the aims are primarily scientific, though many applications will provide demanding tests for the science.

4 What kind of grand challenge is it?

This project arises out of and contributes to two age-old quests: the attempt to understand what we are, and the attempt to make artificial human-like systems, whether entertaining toys, surrogate humans to work in inhospitable environments or intelligent robot helpers for the aged and the infirm.

This is not a grand challenge with a well-defined end-point, like the challenge of getting a human to the moon and back, or the challenge of proving Fermat's last theorem. It is more like the challenge of understanding (as opposed to merely mapping) the human genome or the challenge of curing cancer, both of which require many different problems to be solved, requiring both scientific and technical advances, and potentially producing a wide variety of benefits, but without any definite endpoint. For instance, there are many forms of cancer and some are already successfully treatable, with permanent cures, some are partially treatable and for some there is at present only

palliative treatment. Similarly, understanding the human genome requires advances in many areas of study, including embryology, immune systems, growth and repair mechanisms, the development of brains, the relationships between DNA and species-specific behaviours, etc.

Likewise, although we cannot identify a definite endpoint at which we shall have a complete understanding and working model to demonstrate success of the *Architecture of Brain and Mind* grand challenge, it is the kind of challenge which can be expected to lead to many new scientific and practical advances, provided that people working in the various disciplines concerned understand the global challenge and do their research in the context of contributing to this challenge, instead of being contented with narrowly focused investigation of mechanisms and processes that take no account of how those are related to other mechanisms and processes at a similar level of abstraction and at different levels of abstraction.

Moreover, we can identify some major milestones to be achieved both as tests that we are moving in the required direction and as demonstrations to help others appreciate what has been done and evaluate its interest and importance. Without such a set of milestones, many researchers claiming to be contributing to this project will simply continue what they were going to do anyway. We therefore need to define milestones that require new forms of integration both across implementation levels and within working systems at each level. Defining these milestones will itself be a major task.

4.1 Milestones and backward-chaining

We can do this by trying to specify a kind of integration of varieties of types of functioning at different levels of implementation that we would ideally wish to achieve even if we know that it is likely to take many decades, or possibly even centuries. Working back from that specification by simplifying in various ways we can define a series of less demanding challenges, for instance in 20 years, 10 years, 5 years, 3 years and 1 year where at each stage the tasks will both substantially stretch us beyond what we had previously achieved while clearly moving us in the direction of the next stage.

This is backward-chaining research, in contrast with the more usual forward-chaining research in which people ask what they can or should do next on the basis of what they have already achieved.[9]

Of course one of the consequences of working on some of the intermediate targets will be the discovery that we have not understood some of the problems, leading to re-organisation of the long term roadmap.

It is clear from discussions that have been held on this project as well as familiar differences among researchers in AI, cognitive science, neuroscience, etc. that not everyone will be able to agree on which intermediate steps are achievable, which ones are most worth aiming for, or the order in which things should be done. But that simply means that within the framework of the grand challenge we should allow different intermediate challenges to be pursued in parallel, as long as the researchers frequently meet with others following different paths, and as long as all concerned are agreed on the main features of the more distant end points. Those who are not, can do their research anyway, in a different context: diversity of approaches and goals is part of the essence of science.

[9]The CoSy project, under consideration by the EC, described here http://www.cs.bham.ac.uk/research/cogaff/ adopts very carefully designed, increasingly challenging robotic scenarios as milestones.

5 How will it be done?

Several mutually-informing tasks will be pursued in parallel:

Task 1 Specify, design, and build a succession of computational models of brain function, at various levels of abstraction, designed to support as many as possible of the higher level functions identified in other tasks.

Task 2 Codify typical human capabilities, for instance those shared by young children, including perceptual, motor, communicative, emotional and learning capabilities and use them

- to specify a succession of increasingly ambitious design goals for a fully functioning (partially) human-like system,
- to generate questions for researchers studying humans and other animals which may generate new empirical research leading to new design goals

Task 3 Develop a new theory of the kinds of *architectures* capable of combining all the many information-processing mechanisms operating at different levels of abstraction, and test the theory by designing and implementing a succession of increasingly sophisticated working models meeting the requirements developed in Task 2, each version adding more detail. Analysing what those models can and cannot do and why, will feed back information to the other two tasks.

A possible 15 to 20 year target providing an extremely demanding test of the scientific advances might be demonstration of a robot with some of the general intelligence of a young child, able to learn to navigate a typical home and perform a subset of domestic tasks, including some collaborative and communicative tasks. Unlike current robots it should know what it is doing and why, and be able to discuss alternatives. Linguistic skills should include understanding and discussing simple narratives about things that can happen in its world, and their implications. Manipulative skills could include opening and shutting doors, moving furniture, fetching cups and cutlery from drawers and cupboards, filling a kettle and boiling water, folding clothes and putting them in drawers, carrying objects up and down stairs, including a tray with cups of tea, opening a packet of biscuits and putting the contents on a plate, etc. Social skills might include knowing what another robot or person can or cannot see and using that in deciding what help to give, or how to answer questions. Additional social skills will include understanding motivation, preferences, hopes, fears, etc.

Achieving all this will require major scientific advances in the aforementioned disciplines, including a much richer understanding of the variety of types of information-processing mechanisms and architectures, and could provide the foundation for a variety of practical applications in many industries, in unmanned space exploration, in education, and in the ever-growing problem of caring for disabled or blind persons wishing to lead an active life without being dependent on human helpers. Perhaps many people reading this will welcome such a helper one day.

5.1 Alternative targets

It is not yet clear what sort of stage of human development would make a good target (after much simplification) for a 15 to 20 year project. Some people have argued that we should aim to design

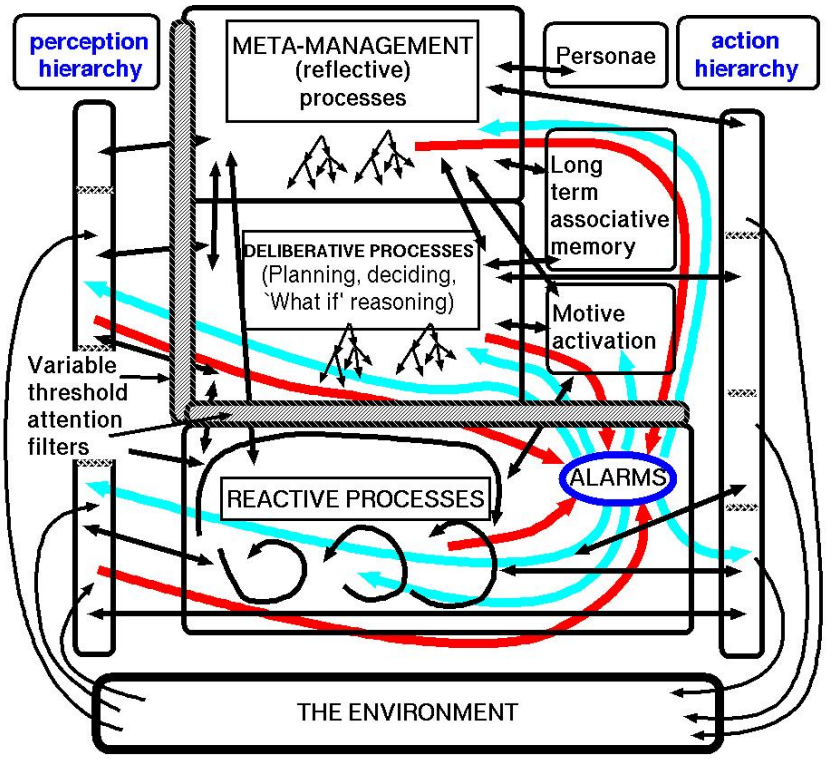

Figure 2: *The H-Cogaff architecture sketch. (More detailed explanation to be added).*

a model of a neonate which can then learn and develop in something like the ways human infants do.

Others argue that human infants are so inscrutable that it is too difficult to know what sorts of information processing their brains are doing, including controlling the development of new brain mechanisms and virtual machine architectures. On this view it may be better to aim for a later stage where there is a huge amount of information about what children can and cannot do and how their abilities change because of the richness of their interaction with their environment, both in action and in linguistic communications. This suggests a strategy of trying to arrive at something like a working model of an older child, perhaps aged 3 to 5 years, though with many simplifications to make the problem tractable. Then we'll have a much clearer idea of what the infant brain might be working towards which could inspire new kinds of research into the earlier stages, especially after more powerful non-invasive brain imaging mechanisms enable us to understand more about the low-level processes.

We do not need to settle this debate: both strategies can be pursued in parallel provided that everyone concerned understands and is clearly contributing to the common long term goal.

A common requirement will be understanding architectures at different levels of abstraction. At present there are many proposals for high level virtual machine architectures, in the literature of AI and Cognitive science, as well as less computationally oriented proposals in the literature of psychology, and psychiatry. Many of the proposals address only a very small subset of the issues that would be need to be addressed in an architecture for a system combing a wide range of human capabilities, even at the level of a young child.

Examples of two closely related very ambitious proposals for human-like virtual architectures at a high level of abstraction can be found in Marvin Minsky's draft book 'The Emotion Machine' available online at his web site[10] and the H-Cogaff architecture being developed by the Cognition and Affect project in Birmingham, shown in Figure 2.

5.2 International collaboration

Several international research programmes, including 'Cognitive Systems' initiatives in Europe and in the USA are now supporting related research: international collaboration is essential for success in such a demanding project.

The four year multi-site 'Integrated Projects' to be supported by the EC's Framework 6 Cognitive Systems initiative will be closely related to the grand challenge aims described here, and some of the smaller projects are also likely to be able to contribute.

TO BE CONTINUED

References

K. Campbell. *Body and Mind*. Macmillan, London, 1970.

David J Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York, Oxford, 1996.

J.J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986. (originally published in 1979).

J. E. Laird, A. Newell, and P. S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

N. Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. The MIT Press, Cambridge, Mass, 1961. 2nd edition.

[10] **http://www.media.mit.edu/~minsky/**