The DAM Symposium: How to Design a Functioning Mind
17-18 April 2000

Introduction: Models of Models of Mind

Aaron Sloman
School of Computer Science, The University of Birmingham
http://www.cs.bham.ac.uk/~axs/

Abstract

'Designing a Mind' abbreviated as 'DAM' is easier to type than the full title of the symposium. Many people are working on architectures of various kinds for intelligent agents. However different objectives, presuppositions, techniques and conceptual frameworks (ontologies) are used by different researchers. These differences together with the fact that many of the words and phrases of ordinary language used to refer to mental phenomena are radically ambiguous, or worse, indeterminate in meaning, leads to much argumentation at cross purposes, misunderstanding, re-invention of wheels (round and square) and fragmentation of the research community. It was hoped that this symposium would bring together many different sorts of researchers, along with a well known novelist with ideas about consciousness, who might, together, achieve something that would not happen while they continued their separate ways. This introduction sets out a conceptual framework which it is hoped will help that communication and integration to occur. That includes explaining some of the existing diversity and conceptual confusion and offering some dimensions for comparing architectures.

1 Introduction

It is now common in Artificial Intelligence and Cognitive Science to think of humans and other animals, and also many intelligent robots and software agents, as having an information processing architecture which includes different layers which operate in parallel, and which, in the case of animals, evolved at different stages. This is not a physical architecture, but something more abstract.

In the early days of AI there was far more talk of algorithms and representations than of architectures, but in recent years it has become clear to many people that we also need to understand how to put various parts (including algorithms and representations) together into a larger working system, and for that an architecture is required.

Some computer scientists still use the word 'architecture' only to refer to the physical or digital electronic architecture of a computer, as was common about 20 or 30 years ago, and still is in courses on computer architectures. However the word can also be used to refer to the architecture of a company, a symphony, a compiler, operating system, a theory or a mind. In particular, it can be used to describe any complex system made of coexisting parts which interact causally in order to serve some complex function or produce some behaviour. The parts may themselves have complex architectures. The system and its parts need not be physical. Nowadays the word often refers to non-physical aspects of computing systems, i.e. *virtual machines*. E.g. an operating system or chess program is a virtual machine with an architecture, though it will need to be implemented in a physical system, usually with a very different architecture.

'Information processing' is another term which has both narrow and broad interpretations: some people restrict it to refer to the kinds of bit-manipulations that computers do. However it can be used to refer to a wide range of phenomena in both discrete and continuous virtual machines of various kinds, including acquiring perceptual information about an environment, storing facts, deriving new consequences, searching a memory or database for answers to questions, creating plans or strategies, generating goals, taking decisions, giving instructions or exercising control. As the last two illustrate, not all information is *factual*: there is also *control* information, including very simple on-off control signals, variations in continuous control parameters, labels for actions to perform, and descriptions of what is to be done.

1.1 Information processing models

Thinking of a brain or mind as an information processing system with an architecture is quite old in philosophy, psychology and neuroscience. The early British empiricist philosophers thought of a mind as made of a collection of 'ideas' (experiences) floating around in a sort of spiritual soup and forming attachments to one another. Kant (1781) proposed a richer architecture with powerful innate elements that enable having experiences and learning from from them to get off the ground, along with mathematical reasoning and other capabilities. About a century ago Freud's division of the mind into 'superego', 'ego' and 'id' (among other things) directed attention to a large subconscious component in the architecture, also implicit in Kant's notion of a schema. Somewhat later Craik (1943) put forward the idea that animals build 'models' of reality in order to explore consequences of actions safely without actually performing them (though it is not clear whether he understood the notion of a model in a virtual machine). Popper (e.g. in his 1976 and earlier works) advocated similar mechanisms allowing our mistaken hypotheses to die instead of us.

Recent work has added more detail, some inspired by neuroscience, some by computational models and some by both. Albus (1981, p.184) depicts MacLean's idea of a 'triune' brain with three layers: a reptilian level and two more recently evolved (old and new mammalian) layers. (This may be insulting to intelligent reptiles.) More recently, AI researchers have been exploring a number of variants, of varying sophistication and plausibility, and varying kinds of control relations between layers. For instance, see Nilsson's (1988, Ch 25) account of triple tower and triple layer models, and various models presented at this symposium, including our own distinction between reactive, deliberative and meta-management layers.

It is also now commonplace to construe many biological processes, including biological evolution and development of embryos as involving acquisition and use of information. Perhaps the biosphere is best construed as an information processing virtual machine driven partly by co-evolutionary interactions.

1.2 Prerequisites for progress

Theories about architectures for minds, brains, or AI systems raise a host of problems. One is that superficially similar architectures may have important differences (some described below) that have not been analysed adequately by researchers. As a result there is no systematic overview of the space of interesting or important architectures, or the different types of requirements which architectures may be required to satisfy, against which they can be evaluated. In short there are no adequate surveys of 'design space' and 'niche space' and their relationships. See Sloman (1994, 1998b).

A worse problem is that there is considerable terminological confusion, obscured by the confidence with which people use words and phrases referring to mental states and processes, including, for example, 'belief', 'desire', 'intention', 'consciousness', 'learning', 'emotion', 'personality', 'understanding', and many others.

AI researchers who blithely use mentalistic labels to describe various mechanisms on the basis of shallow analogies were berated long ago by McDermott (1981). However the habit does not die easily.

Moreover, a social psychologist interested in human relations is likely to define 'emotion' so as to cover the phenomena associated with social relationships such as embarrassment, attachments, guilt, pride, loyalty, etc., whereas a brain scientist studying rodents may define the word so that it refers to the brain processes and observable behaviours found in such animals. Other foci of interest lead to yet more definitions of 'emotion' and there are dozens of them in the psychological and philosophical literature. By taking a broader view than any of their proponents, we should be able explain how to accommodate all of these definitions (at least those related to real phenomena) in the same framework in the same general framework.

1.3 Architecture-based concepts

The task of getting a clear overview of the variety of information processing architectures and the problems of clarifying our confused concepts are closely connected.

That is because each architecture supports a collection of capabilities, states and processes, and different clusters of such capabilities and the states and processes define different concepts. For example an operating system that does not support multi-processing cannot support the distinction between thrashing and not thrashing nor does it make sense to ask about its interrupt priority levels. Likewise an architecture for an animal or robot supports a family of mental concepts and different architectures support different families.

Thus we need to be clear about the architectural presuppositions of our concepts. Otherwise, different researchers will focus attention on different aspects of reality, and adopt definitions suited to their interests, not realising that they are ignoring other equally important phenomena, like the proverbial group of blind people each trying to describe an elephant on the basis of what they individually can feel.

It is not hard to convince a blind man that he is in contact with only a small region of a large structure. It is much harder to convince people producing theories of mind that they are attending to a tiny part of a huge system. Psychologists have produced dozens of distinct definitions of 'emotion', and instead of taking this as a clue that there is a range of diverse phenomena which should be given different labels, they often argue about which definition is 'correct'. Our own analysis of various sorts of human emotions has begun to show how in a suitably rich architecture, several different types of processes can occur which correspond to what we sometimes call emotions, which we now distinguish as primary, secondary and tertiary emotions, extending the classification of Damasio and others. See Damasio (1994); Picard (1997); Sloman (1998a, 2000); Sloman and Logan (2000).

2 Deceptive clarity

Evolution has produced brains which, at least in humans, give their owners some information about their own internal processing. This information is deceptively compelling, and often thought to be incapable of being erroneous because it is so direct. We seem to have direct access to our thoughts, decisions, desires, emotions and, above all our own consciousness. This familiarity leads many people to think they know exactly what they are talking about when they engage in debates about the nature of mind, and propose theories about consciousness, experience, awareness, the 'first-person viewpoint', and so on.

However, the diversity of opinions about the nature of the phenomena, especially the widely differing definitions offered by various psychologists, cognitive scientists, brain scientists, AI theorists and philosophers of terms like 'emotion' and 'consciousness', casts serious doubt on the assumption that we all know what we are referring to.

2.1 Two sources of confusion

The confusion has several roots, one of which is the hidden complexity and diversity of the phenomena: the architectural presuppositions of human mentality are extraordinarily complex, and still far from being understood. Moreover there are differences not only between human beings at different stages of development or when suffering from various kinds of damage or disease, but also between humans and different sorts of animals and artefacts. So if mental concepts are inherently architecture-relative the study of mind will require many families of concepts to describe all the phenomena adequately, unlike the study of the physical world. Of course different concepts are required for different levels in the physical ontology, e.g. sub-atomic physics, chemistry, astrophysics, geology, etc. In contrast, concepts of mind involve both differences of levels and differences of architectures at all levels.

Another source of confusion is a common type of philosophical error, namely believing that we have a clear understanding of concepts just because they refer to phenomena that we experience directly. This is as mistaken as thinking we fully understand what simultaneity is simply because we have direct experience of seeing a flash and hearing a bang simultaneously. Einstein taught us otherwise.

From the fact that we can recognise some instances and non-instances of a concept it does not follow that we know what is meant *in general* by saying that something is or is not an instance. There are endless debates about which animals have consciousness, whether machines can be conscious, whether unborn infants have experiences, or whether certain seriously brain-damaged humans still have minds. Our disagreement even over what counts as relevant evidence, is a symptom that our concepts of mentality are far more confused than we realise.

There is no point attempting to resolve such questions by empirical research when we cannot agree on which evidence is relevant. Does wincing behaviour in a foetus prove that it feels pain and is therefore conscious, or is it a mere physiological reaction? How can we decide? Does the presence of a particular type of neural structure prove that the foetus (or some other animal) is conscious, or is the link between physical mechanisms and consciousness too tenuous for any such proof to be possible, as many philosophers have argued?

We can explain why there is so much confusion and disagreement by exposing the hidden complexity of the presuppositions of our ordinary concepts, the diversity of the phenomena referred to, and the indeterminateness of most of our 'cluster' concepts.

2.2 Cluster concepts

Many concepts, besides being architecture-based, are 'cluster concepts', referring to ill-defined clusters of capabilities and features of individuals. If an architecture supports capabilities of types C1, ...Ck and produces processes with features F1, ...Fn, then different combinations of those capabilities and features can define a wide variety of states and processes. But our pre-theoretical cluster concepts lack that kind of precision. For a given mental concept M there may be some combinations of Cs and Fs that definitely imply presence of M, and others which definitely imply absence of M, but there need not be any well-defined boundary between instances of M and non-instances. That is shown by the intense debates about intermediate cases.

This does not mean that there is a fuzzy or probabilistic boundary. Fuzzy boundaries sometimes occur where there is smooth variation and a probabilistic classifier is at work. With cluster concepts there can be clear cases at extremes and total indeterminacy as regards a wide range of intermediate cases, because there has never been any need, nor any basis, for separating out the intermediate cases.

Making all this clear will show how we can define different families of more precise concepts related to the capabilities supported by different architectures. Which definitions are *correct* is a pointless question, like asking whether mathematicians are 'correct' in defining 'ellipse' so as to include circles. Wheel-makers and mathematicians have different concerns.

2.3 Refining and extending concepts

When we have a clear view of the space of architectures that are of interest (including architectures for human-like systems, for other animals, for various kinds of robots and for various sorts of software agents) we can then consider the families of concepts generated by each type of architecture. We can expect some architectures to support some of our mental concepts (in simplified forms) e.g. 'sensing', but not necessarily all of our notions of 'pain', 'emotion', 'intelligence', 'consciousness', etc.

For instance, an insect has some sort of awareness of its environment even if it has nothing like full human consciousness, e.g. if it is not aware that it is aware of its environment. Precisely which sort of awareness

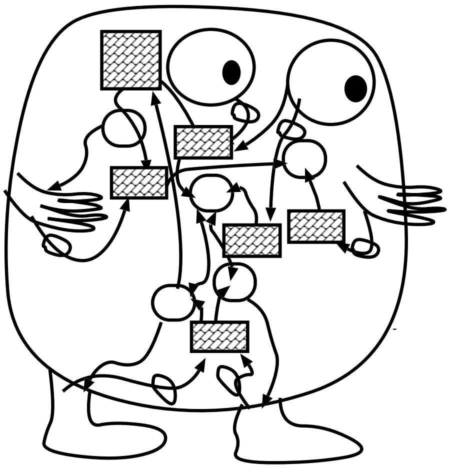

Figure 1: A possible unstructured architecture
In principle, an architecture might be a completely unstructured mess which we could never hope to understand. This is how some view products of evolution. Alternatively evolution, like human designers, may be incapable of producing very complex successful designs unless they have a high degree of structure and modularity, which can provide a principled basis for defining concepts of types of states and processes that can occur.

it has cannot be answered without knowing about its information processing architecture.

Similarly it may be acceptable to use simplified forms of our ordinary concepts in describing some existing AI systems, even though none of them comes close to matching typical human mentality. And if we had a clear idea of the information processing architecture of a foetus at different stages of development, then for each stage we could specify concepts of pain, or awareness that are relevant. However, we should not assume that all concepts applicable to adult humans will be relevant. For instance, it is almost certain that a foetus, or even a new-born infant is not yet capable of being puzzled about the relationship between its mental states and its body or wondering whether a good deity would allow pain to exist. It is possible that a new born infant lacks an architecture capable of supporting wondering about anything.

3 What sorts of architectures?

We know so little about possible information processing mechanisms and architectures (especially the extraordinarily powerful visual mechanisms implemented in animal brains) that it is premature to hope for a complete survey of types of architectures and their capabilities. It could turn out, as some have claimed, that any information-processing architecture produced by millions of years of evolution is bound to be far too messy and unstructured for us to understand as engineers, scientists or philosophers (Figure 1).

Alternatively, it may turn out that evolution, like human designers must use principles of modularity and re-usability in order to achieve a robust and effective collection of architectures, such as we find in many kinds of animals. Figures 2(a) and (b) indicate more structured and modular architectures, combining a three-fold division between perception, central processing, and action, and three levels of processing, with and without a global 'alarm' mechanism. However, such diagrams can be misleading partly because they convey very different designs to different researchers. A frequent confusion is between diagrams indicating state-transitions (flow-charts) and diagrams indicating persisting, interacting components of an architecture. In the former an arrow represents a possible change of state. In the latter it represents flow of information between components. My diagrams are of the latter kind.

To help us understand what to look for in naturally occurring architectures, it may be useful to attempt a preliminary overview of some features of architectures that have already been proposed or implemented. We can then begin to understand the trade-offs between various options and that should help us to understand the evolutionary pressures that shaped our minds.

3.1 Layered architectures

Researchers on architectures often propose a collection of layers. The idea of hierarchic control systems is very old both in connection with analog feedback control and more recently in AI systems. There are many proposals for architectures with three or more layers, including those described by Albus and Nilsson mentioned previously, the subsumption architecture of Brooks (1991), the ideas in Johnson-Laird's discussion (1993) of consciousness as depending on a high level 'operating system', the multi-level architecture proposed for story understanding in Okada and Endo (1992), Minsky's notion of A, B and C brains in section 6.4 of Minsky (1987) and also in several of the papers at this conference.

3.2 Dimensions of architectural variation

On closer inspection, the layering in multi-level architectures means different things to different researchers. There seem to be several orthogonal distinctions at work, which, at present, I can only classify very crudely.
1. Concurrently active vs pipelined layers
In Albus (1981) and some of what Nilsson (1998) writes, the layers have a sequential processing function: sensory information comes in (e.g. on the 'left') via sensors to the bottom layer, gets abstracted as it goes up through higher layers, then near the top some decision is taken, and then control information flows down through the layers and out to the motors (on the other side). I call this an "Omega" architecture because the pattern of information flow is shaped like an Ω. Many AI models have this style. The

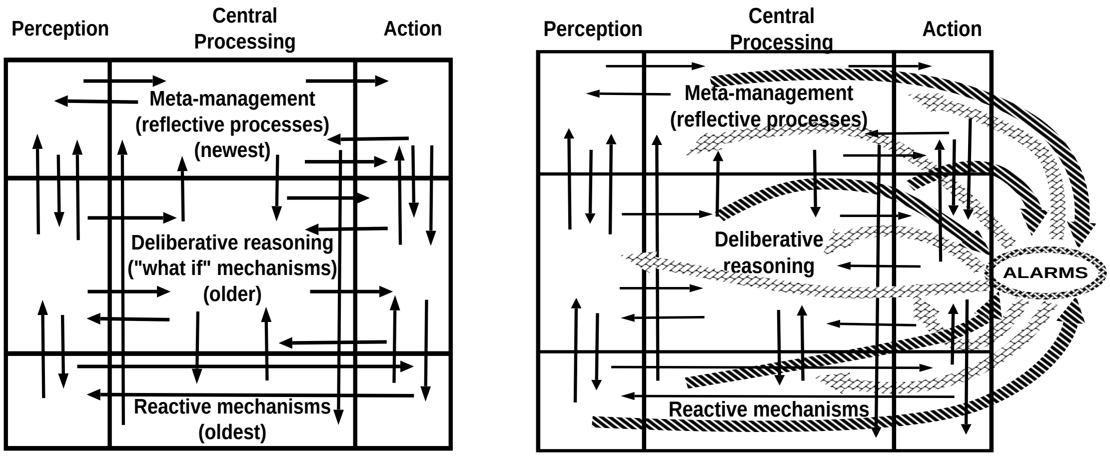

Figure 2: (a) (b)

Nilsson distinguishes 'triple tower' models, with information flowing (mainly) in through a perceptual tower to a central processing tower, then out to a motor tower, and 'triple layer' models where different layers perform different functions. Depending on processing speeds in these mechanisms there may also be a need for a fast global 'alarm' mechanism. Figure (a) serves as a mnemonic indicating the triple tower and triple layer views superimposed, where the various components in the boxes will have functions defined by their relationships with other parts of the system. In (b) a global alarm system is indicated, receiving inputs from all the main components of the system and capable of sending control signals to all the components. Since such alarm systems need to operate quickly when there are impending dangers or short-lived opportunities, they cannot make use of elaborate inferencing mechanisms, and must be pattern based. Global alarm mechanisms are likely therefore to make mistakes at times, though they may be trainable.

enhanced version of Norman and Shallice's "contention scheduling" model, described in Glasspool's contribution to this symposium, is a variant of the Omega schema in which the upward information flow activates a collection of competing schemata where winners are selected by a high level mechanism for controlling attention.

An alternative is an architecture where the different layers are all concurrently active, with various kinds of control and other information constantly flowing within and between them in both directions, as in figure 2 and the 'Cogaff' architecture in 3.

2. Dominance hierarchies vs functional differentiation

A second distinction concerns whether higher levels *dominate* lower levels or merely attempt to control them, not always successfully and sometimes with the direction of control reversed. In the subsumption model (Brooks 1991) higher levels not only deal with more abstract state specifications, goals and strategies, but also completely dominate lower levels. I.e. they can turn lower level behaviour off, speed it up, slow it down, modulate it in other ways, etc. This conforms to the standard idea of hierarchical control in engineering.

By contrast, in a non-subsumptive layered architecture (figures 2 and 3) the 'higher' levels manipulate more sophisticated and abstract information, but do not necessarily dominate the lower levels, although they may sometimes attempt to do so. Higher levels may be able partially to control the lower levels but sometimes they lose control, either via alarm mechanisms or because other influences divert attention, such as sensory input with high salience (loud noises, bright flashes) or newly generated motives with high 'insistence' (e.g. hunger, sitting on a hard chair, etc.). In such a model the *majority* of lower level reactive mechanisms cannot be directly controlled by the deliberative and metamanagement layers, especially those concerned with controlling bodily functions. Some training may be possible, however.

3. Direct control vs trainability

In some layered systems it is assumed that higher levels can directly control lower levels. A separate form of control which is not 'immediate' is re-training. It is clear that in humans higher levels can sometimes retrain lower levels even when they can't directly control them.

For instance, repeated performance of certain sequences of actions carefully controlled by the deliberative layer can cause a reactive layer to develop new chained condition-action behaviour sequences, which can later run without higher level supervision. Fluent readers, skilled athletes, musical sight-readers, all make use of this. (The nature of the boundary between central mechanisms and action control mechanisms is relevant here.)

4. Different kinds of processing vs different control functions

On some models different layers all use the same kinds of processing mechanisms (e.g. reactive behaviours) but perform different functions, e.g. because they operate at different levels of abstraction. In other models there are different kinds of processing as well as different functional roles.

For instance, Figures 2 and 3 present a lowest level that is purely reactive, whereas the second and third levels can do deliberative, 'what if', reasoning, using mechanisms able to represent possible future actions and consequences of actions, categorise them, evaluate them,

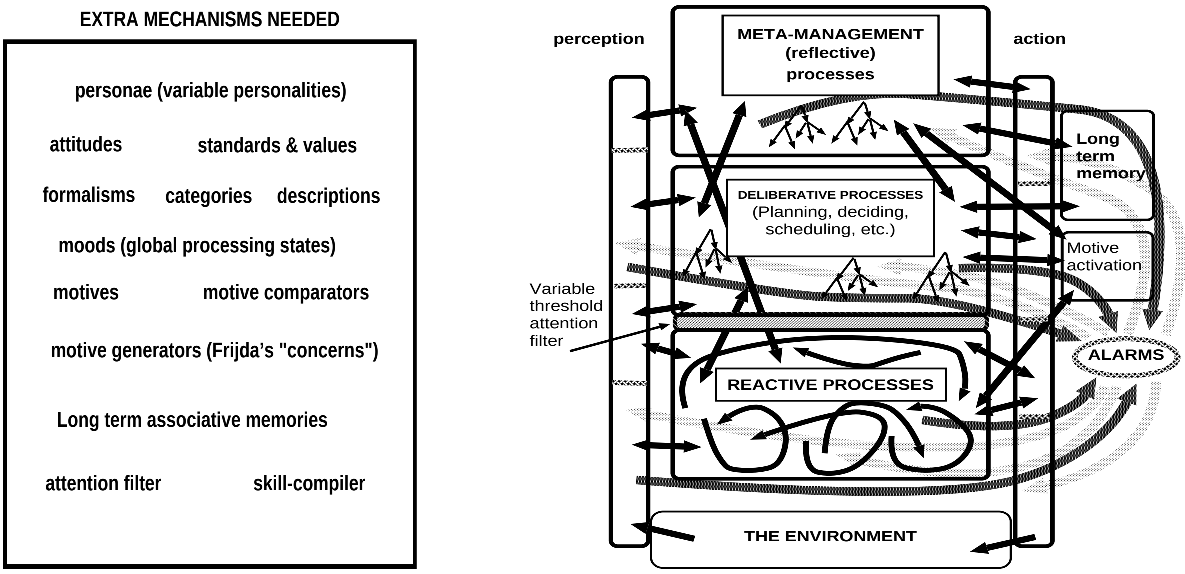

Figure 3: (a) (b)

The Birmingham Cogaff Architecture

We have been exploring ideas based on the collection of mechanisms depicted in Figure 2(b) enhanced with additional components required to make everything work. In (a) we list some additional components required to support processing of motives, 'what if' reasoning capabilities in the deliberative layer, and aspects of self-control. It is conjectured that there is a store of different, culturally influenced, 'personae' which take control of the top layer at different times, e.g. when a person is at home with family, when driving a car, when interacting with subordinates in the office, in the pub with friends, etc. In (b) relations between some of the components are shown along with a global alarm system, receiving inputs from everywhere and sending interrupt and redirection signals everywhere. It also shows a variable-threshold interrupt filter, which partly protects resource-limited deliberative and reflective processes from excessive diversion and redirection. The filter should be thought of as 'wrapped around' the higher levels, with a dynamically varying penetration threshold, dependent, for instance, on the urgency and importance of current tasks.

and make selections. This is not how reactive systems behave. Traditional AI planning systems can do this, and similar mechanisms are able to explain past events, do mathematical reasoning, or do general reasoning about counterfactual conditionals. However, it is possible, indeed likely, that the deliberative mechanisms which go beyond reactive mechanisms in explicitly representing alternative actions prior to selection are themselves *implemented* in reactive mechanisms, which can operate on structures in a temporary workspace.

Reactive mechanisms may be implemented in various kinds of lower level mechanisms, including chemical, neural and symbolic information processing engines, and it is possible that the reliance on these is different at different levels in the architecture. Some kinds of high level global control may use chemical mechanisms which would be too slow and unstructured for intricate problem solving.

Some have argued that human capabilities require quantum mechanisms though I have never seen a convincing account of how they could explain any detailed mental phenomena.

5. Where are springs of action

A fifth distinction concerns whether new 'intrinsic' motives (which are not sub-goals generated in a planning process) all come from a single layer or whether they can originate in any layer. In one variant of the Omega model, information flows up the layers and triggers motivational mechanisms at the top. In other models, processes anywhere in the system may include motive generators, for instance physiological monitors in the reactive layer. The motives they generate may be handled entirely by reactive goal-directed behaviours, or they may need to be transferred to the deliberative layer for evaluation, adoption or rejection, and possibly planning.

6. Handling competing motives

Not all motives will be mutually consistent, so there has to be some way of dealing with conflicts. Architectures differ regarding the locus of such conflict resolution and the mechanisms deployed.

For instance, in some forms of contention-scheduling models, schemata form coalitions and oppositions on the basis of fixed excitatory and inhibitory links in a network, and then some kind of numerical summation leads to selection, which is always done at the same level in the hierarchy. In other models the detection of conflicts might use symbolic reasoning, and the resolution might be done at different levels for different sorts of conflicts.

For instance the decision whether to help granny or go to the marvellous concert might be handled in one part of the system, and the decision whether to continue uttering the current unfinished sentence or to stop and take a breath another way, and the decision to use placatory or abusive vocabulary when addressing some who has angered you might be handled by yet another part of the system.

7. Perceptual to central connections

Architectures with perceptual components differ in the relationships between modes of processing in perceptual modules and more central layers. E.g. is the perceptual processing itself layered, producing different levels of perceptual information to feed into different central layers, or is there a fixed entry level into the central mechanisms, after which the information may or may not be passed up a hierarchy, as in the Omega model. The latter might be described as the 'peephole' model of perception the former the 'multi-window' model of perception.

In 'peephole' perceptual systems, the sensory mechanisms (simple transducers or more complex sensory analysers) produce information about the environment and direct it all to some component of the central architecture. That may trigger processes which affect other parts.

In Figures 2 and 3 it is suggested that the perceptual processes are themselves layered, handling different levels of abstraction concurrently, with a mixture of top-down and bottom up processing, and with different routes into different parts of the central system. For instance deliberative mechanisms may need perceptual information chunked at a fairly high level of abstraction, whereas fine control of movement may require precise and continuously varying input into the reactive system. Differential effects of different kinds of brain damage seem to support the multi-window multi-pathway model, which can also be defended on engineering grounds.

8. Central to motor connections

An analogous distinction concerns the relationship between central and motor processing. Just as there is what I called 'multi-window' perception and 'peephole' perception, so too with action. At one extreme there is only a 'narrow' channel linking the motor system only with the lowest level central mechanism, as in the Omega model: there are motors and they all get signals directly from one part of the central mechanism (analogous to 'peephole' perception). At another extreme there can be a layered, hierarchical motor control system where control information of different sorts comes in directly at different levels, from different layers in the central system.

Humans seem to have motor systems with complex hierarchical skills, and probably also many other animals.

In some proposed architectures (e.g. Albus (1981)) this hierarchical organisation of action is acknowledged, but instead of the action hierarchy being a separate 'tower' (in Nilsson's terminology) communicating with several central processing layers it is folded in to the central control hierarchy. Of course, the two models could describe equivalent systems, but it may sometimes be more useful to think of the central system and the action systems as both having hierarchic organisation. This may help us understand how the whole system evolved in humans and other animals and the increased modularity may help with design tasks. However that is still only a conjecture. Similar comments are applicable to different architectures for perception.

9. Emergence vs 'boxes'

One of the notable features of recent AI literature is the proliferation of architecture diagrams in which there is a special box labelled 'emotions'. Contrast Figures 2 and 3, where there is no specific component whose function is to produce emotions, and instead emotions are explained as emergent properties of interactions between components which are there for other reasons, such as alarm mechanisms and mechanisms for diverting attention (which can happen without any emotion being generated). Elsewhere I have shown how at least three different classes of emotions (primary, secondary and tertiary emotions) emerge in the three layer 'Cogaff' architecture. (This may be compared with the emergence of 'thrashing' in a multi-processing architecture. The thrashing is a result of heavy load and interactions between mechanisms for paging, swapping and allocating resources fairly.)

The problem may be partly terminological: e.g. some theorists write as if all motives are emotions. Then a component that can generate motives may be described as an 'emotion generator' by one person and as a 'motive generator' by another. Separating them accords better with ordinary usage, since it is possible to have motives and desires without being at all emotional, e.g. when hungry. This is just one of many areas where we need far greater conceptual clarity, which may come in part from further study of varieties of architectures their properties, and the state transitions they support.

There are probably many cases where it is not clear whether some capability needs to be a component of the architecture, or an emergent feature of interactions between components. The attention filter in Figure 3(b) is an example. Instead of a special filtering mechanism, the effects of filtering may be produced by interactions between competing components. The first alternative may be easier to implement and control. The second may be more flexible and general. There are many design tradeoffs still to be analysed.

10. Dependence on language

Some models postulate a close link between high level internal processes and an external language. For instance, it is often suggested (Rolls 1998) that mechanisms analogous to meta-management could not exist without a public language used by social organisms, and in some of Dennett's writings consciousness is explained as a kind of 'talking to oneself'.

A contrary view is that internal mechanisms and formalisms for deliberation and high level self-evaluation and control were necessary pre-cursors to the development of human language as we know it.

The truth is probably somewhere in between, with an interplay between the development of internal facilitating information processing mechanisms and social processes which then influence and enhance those mechanisms, for instance by allowing a culture to affect the development in individuals of categories for internal processes of self-evaluation. (Freud's 'super-ego'). However, it appears from the capabilities of many animals without what we call language, that very rich and complex information processing mechanisms evolved long before external human-like languages, and probably still underpin them. We could extend the word 'language' to refer to forms of internal representation and say that the use of language to think with is prior to its use in external communication.

11. Purely internal vs partly external implementation

A more subtle distinction concerns how far the implementation of an organism or intelligent artefact depends entirely on the internal mechanisms and how far the implementation is shared with the environment. The development in the 70's of 'compliant wrists' for robots, which made it far easier, for example, to program the ability to push a cylinder into a tightly fitting hole, illustrated the advantage in some cases of off-loading information processing into mechanical interactions. Trail-blazing and the design of ergonomically effective tools and furniture are other examples.

From a philosophical viewpoint a more interesting case is the ability to refer to a spatially located individual unambiguously. As explained long ago in Strawson (1959), whatever is *within* an individual cannot *suffice* to determine that some internal representation or thought refers to the Eiffel tower, as opposed to an exactly similar object on a 'twin earth'. Instead the referential capability depends in part on the agent's causal and spatial relationships to the thing referred to. So attempting to implement *all* aspects of mental functioning entirely within a brain or robot is futile: there is always a subtle residue that depends on external relations. (In referring to parts of oneself, or parts of one's own virtual machine the problem is solved internally, as explained in Sloman (1985, 1987).)

12. Self-bootstrapped ontologies

I have been arguing that when we have specified an architecture we shall understand what sorts of processes can occur in it, and will be able to define an appropriate set of concepts for describing its 'mental' states.

However, some learning mechanisms can develop their own ways of clustering phenomena according to what they have been exposed to and various other things, such as rewards and punishments. If a system with the kind of meta-management layer depicted in the Cogaff architecture uses that ability on itself, it may develop a collection of concepts for categorising its own internal states and processes that nobody else can understand because nobody else has been through that particular history of learning processes. The role those concepts play in subsequent internal processing may exacerbate the uniqueness, complexity and idiosyncratic character of its internal processing.

For systems with that degree of sophistication and reflective capability, scientific understanding of what is going on within it may forever be limited to very coarse-grained categorisations and generalisations. This could be as true of robots as of humans, or bats Nagel (1981).

4 Human-like architectures

I have tried to bring out some of the design options that need to be faced when trying to explain the architecture of a human mind. When we understand what that architecture is, we can use it to define collections of concepts that will be useful for describing human mental states and processes, though we can expect to do that only to a certain degree of approximation for the reasons in the previous paragraph. However that may suffice to provide useful clarifications of many of our familiar concepts of mind, such as 'desire', 'moods', 'emotion' and 'awareness'.

In particular, so many types of change are possible in such complex system that we can expect to find our ordinary concepts of 'learning' and 'development' drowning in a sea of more precise architecture-based concepts.

We may also be in a better position to understand how, after a certain stage of evolution, the architecture supported new types of interaction and the development of a culture, for instance if the meta-management layer, which monitors, categorises, evaluates and to some extent controls or redirects other parts of the system, absorbs many of its categories and its strategies from the culture. It seems that in humans the meta-management layer is not a fixed system: not only does it develop from very limited capabilities in infancy, but even in a normal adult it is as if there are different personalities "in charge" at different times and in different contexts (e.g. at home with the family, driving a car, in the office, at the pub with mates).

This suggests new ways of studying how a society or culture exerts subtle and powerful influences on individuals through the meta-management processes. The existence of the third layer does not presuppose the existence of an external human language (e.g. chimpanzees may have some reflective capabilities), though it does presuppose the availability of some internal formalism, as do the reactive and deliberative layers.

When an external language develops, *one* of its functions may be to provide the categories and values to be used by individuals in judging their own mental processes (e.g. as selfish, or sinful, or clever, etc.) This would be a powerful form of social control, far more powerful than mechanisms for behavioural imitation, for instance. It might have evolved precisely because it allows what has been learnt by a culture to be transmitted to later generations far more rapidly than if a genome had to be modified. However, even without this social role the third layer would be useful to individuals, and that might have been a requirement for its original emergence in evolution.

We can also hope to clarify more technical concepts. The common reference to "executive function" by psychologists and brain scientists seems to conflate aspects of the deliberative layer and aspects of the meta-management layer. That they are different is shown by the existence of AI systems with sophisticated planning and problem solving and plan-execution capabilities without meta-management (reflective) capabilities. A symptom would be a planner that doesn't notice an obvious type of redundancy in the plan it produces, or subtle looping behaviour.

One consequence of having the third layer is the ability to attend to and reflect on one's own mental states, which could cause intelligent robots to discover qualia, and wonder whether humans have them.

All this should provide much food for thought for AI researchers working on multi agent systems, as well as philosophers, brain scientists, social scientists and biologists studying evolution. I hope the DAM symposium makes some useful contribution to the clarification of these ideas.

Acknowledgements and Notes

This work has benefited from discussions with Brian Logan, Marvin Minsky, and many colleagues and students in the Cognition and Affect project at The University of Birmingham. The work is supported by the Leverhulme Trust. Our papers can be found at
http://www.cs.bham.ac.uk/research/cogaff/
and our tools at
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

References

James S. Albus. *Brains, Behaviour and Robotics*. Byte Books, McGraw Hill, Peterborough, N.H., 1981.

R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

Kenneth Craik. *The Nature of Explanation*. Cambridge University Press, London, New York, 1943.

Antonio R Damasio. *Descartes' Error, Emotion Reason and the Human Brain*. Grosset/Putnam Books, 1994.

P.N. Johnson-Laird. *The Computer and the Mind: An Introduction to Cognitive Science*. Fontana Press, London, 1993. (Second edn.).

Immanuel Kant. *Critique of Pure Reason*. Macmillan, London, 1781. Translated (1929) by Norman Kemp Smith.

D. McDermott. Artificial intelligence meets natural stupidity. In John Haugeland, editor, *Mind Design*. MIT Press, Cambridge, MA, 1981.

M. L. Minsky. *The Society of Mind*. William Heinemann Ltd., London, 1987.

Thomas Nagel. What is it like to be a bat. In D.R. Hofstadter and D.C.Dennett, editors, *The mind's I: Fantasies and Reflections on Self and Soul*, pages 391–403. Penguin Books, 1981.

Nils J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, San Francisco, 1998.

N. Okada and T. Endo. Story generation based on dynamics of the mind. *Computational Intelligence*, 8: 123–160, 1992. 1.

Rosalind Picard. *Affective Computing*. MIT Press, Cambridge, Mass, London, England, 1997.

Karl Popper. *Unended Quest*. Fontana/Collins, Glasgow, 1976.

Edmund T. Rolls. *The Brain and Emotion*. Oxford University Press, Oxford, 1998.

A. Sloman. What enables a machine to understand? In *Proc 9th IJCAI*, pages 995–1001, Los Angeles, 1985.

A. Sloman. Reference without causal links. In J.B.H. du Boulay, D.Hogg, and L.Steels, editors, *Advances in Artificial Intelligence - II*, pages 369–381. North Holland, Dordrecht, 1987.

A. Sloman. Explorations in design space. In A.G. Cohn, editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester, 1994. John Wiley.

A. Sloman. Damasio, Descartes, alarms and meta-management. In *Proceedings International Conference on Systems, Man, and Cybernetics (SMC98)*, pages 2652–7. IEEE, 1998a.

A. Sloman. The "semantics" of evolution: Trajectories and trade-offs in design space and niche space. In Helder Coelho, editor, *Progress in Artificial Intelligence, 6th Iberoamerican Conference on AI (IBERAMIA)*, pages 27–38. Springer, Lecture Notes in Artificial Intelligence, Lisbon, October 1998b.

Aaron Sloman. Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Kerstin Dautenhahn, editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam, 2000.

Aaron Sloman and Brian Logan. Evolvable architectures for human-like minds. In *Proceedings 13th Toyota Conference, on Affective Minds Shizuoka, Japan, Nov-Dec 1999*. Elsevier, 2000.

P. F. Strawson. *Individuals: An essay in descriptive metaphysics*. Methuen, London, 1959.