To appear in M Wooldridge and A Rao (Eds) *Foundations of Rational Agency*, **Kluwer Academic Publishers, 1997**
Expanded version of invited talk at Cognitive Modeling Workshop, AAAI96, Portland Oregon, Aug 1996.

What sort of architecture is required for a human-like agent?

Aaron Sloman
School of Computer Science & Cognitive Science Research Centre
The University of Birmingham, B15 2TT, UK
A.Sloman@cs.bham.ac.uk, http://www.cs.bham.ac.uk/~axs

Abstract

This paper is about how to give human-like powers to complete agents. For this the most important design choice concerns the overall architecture. Questions regarding detailed mechanisms, forms of representations, inference capabilities, knowledge etc. are best addressed in the context of a global architecture in which different design decisions need to be linked. Such a design would assemble various kinds of functionality into a complete coherent working system, in which there are many concurrent, partly independent, partly mutually supportive, partly potentially incompatible processes, addressing a multitude of issues on different time scales, including asynchronous, concurrent, motive generators. Designing human like agents is part of the more general problem of understanding design space, niche space and their interrelations, for, in the abstract, there is no one optimal design, as biological diversity on earth shows.

1 Introduction

A complete functioning agent, whether biological, or simulated in software, or implemented in the form of a robot, needs an integrated collection of diverse but interrelated capabilities, i.e. an architecture. At present, most work in AI and Cognitive Science addresses only components of such an architecture (e.g. vision, speech understanding, concept formation, rule learning, planning, motor control, etc.) or mechanisms and forms of representation and inference (logic engines, condition-action rules, neural nets, genetic algorithms) which might be used by many components. While such studies can make useful contributions it is important to ask, from time to time, how everything can be put together, and that requires the study of architectures.

Analysing possible architectures is closely related to the task of defining an ontology for mental objects, states and processes (percepts, beliefs, desires, attitudes, intentions, moods, emotions, character, inferences, learning, etc.). Ideas about the ontology can help to guide design choices. However, exploring an architecture can reveal unexpected features of the ontology it is capable of supporting, and that can feed back into new ideas about ontologies and design requirements. So the processes of theorising, designing, implementing and experimenting are related in a cyclic fashion.

At present I do not think we know much about the space of possible architectures, and our ideas regarding the ontology to be supported by such an architecture are still very primitive (having advanced little beyond folk psychology, though that's as good a starting place as any). So we are not yet in a position to choose one architecture, or even a sub-class. So all such work must remain exploratory and speculative for the time being, including the work reported here.

2 What is an architecture?

What do I mean by “architecture"? A fully functioning system has architectures at different levels of abstraction, corresponding to different implementation layers, e.g. there is the architecture of an underlying physical mechanism (Turing machine, von Neumann machine, dataflow machine, neural net, chemical control mechanism, etc.), the architecture of a complex algorithm (e.g. a parsing algorithm which has components that handle different types of sub-structure in the input), the architecture of an integrated collection of concurrent software modules (e.g. the architecture of an operating system, or the architecture of a factory control system). When computer scientists talk about architecture they often mean to refer to the structure of the lowest level physical mechanism. There is a more important notion of architecture for our purposes, which is closer to what we mean by the architecture of a building, or a large organisation. This refers to the large scale functional decomposition: it is the concept of architecture that might be used by a software engineer, or systems analyst.

Besides differences in levels of abstraction or implementation, there are differences in types of functionality. A human-like agent needs to be able to perform a large and diverse collection of tasks, both externally (finding and consuming food, avoiding predators, building shelters, making tools, finding mates, etc.) and internally (interpreting sensory data, generating motives, evaluating motives, selecting motives, creating plans, storing information for future use, making inferences from new or old information, detecting inconsistencies, monitoring plan execution, monitoring various kinds of internal processing, noticing resemblances, creating new concepts and theories, discovering new rules, noticing new possibilities, etc.).

At present we do not know much about the range of internal tasks performed by the human architecture since neither observation of behaviour, nor introspection nor neurophysiological studies can give direct insight into most of what is going on in abstract virtual machines (for reasons indicated below). Nevertheless we can start our exploration from our best current hunches gleaned from all these sources.

3 There is no unique design for intelligence

Even if the list of internal capabilities given above is a good start, we must not assume that all intelligent agents will have the same collection. Different kinds of agents may have different subsets. Even among humans there is enormous diversity, especially if we consider extreme cases, such as Newton, Mozart, and idiot savants. Within an individual the collection of capabilities is not fixed either, as is clear both from observation of young children and studies of aging.

Thus we should not assume that an intelligent agent has a fixed architecture: part of the processes of learning and development may include changes to the architecture, for instance development of major new collections of capabilities and development of new links between old capabilities. Some individuals seem to go on developing and extending their architectures longer than others. It may turn out that one of the *most* important features of a human architecture, a source of much of its power, is the potential for self modification and the consequential diversification within a cooperating community.

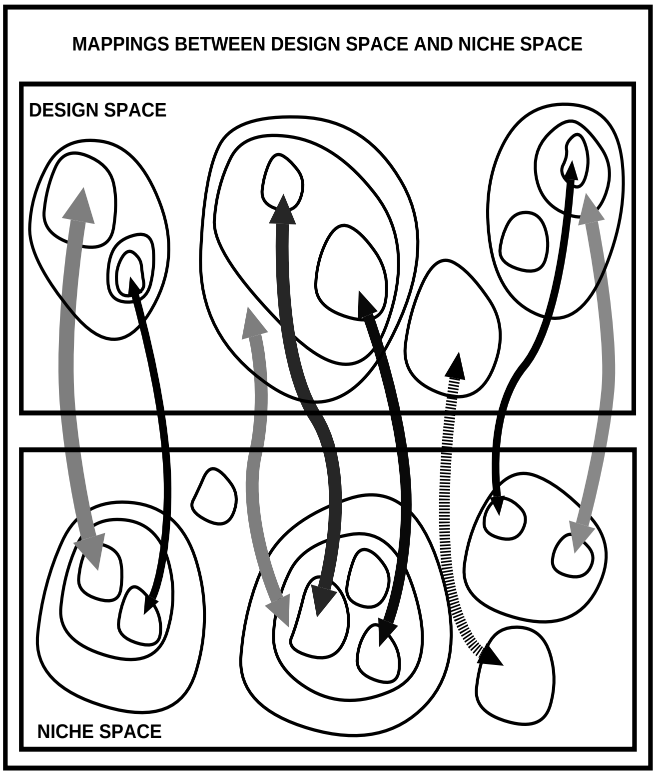

Figure 1: **Mappings between design space and niche space**

4 Design space and niche space

For any collection of capabilities (i.e. for each set of requirements for a design) we can consider the designs that might implement such capabilities. In general there will not be unique design solutions. I have summarised this in [10, 12, 14] by suggesting that we need to explore a space of possible designs for behaving systems (design space) and a space of possible sets of requirements (niche space) and the mappings between the two. It is not to be expected that there is any one "right" architecture. As biological diversity demonstrates, many different architectures can be successful, and in different ways. There are different "niches" (sets of requirements and constraints) for which architectures can be evaluated and compared, and such evaluations will not generally yield a Yes/No decision, but rather an analysis of trade-offs, often involving several dimensions of comparison. This comment does not imply that the spaces are smooth continua without any sharp boundaries: on the contrary, both are likely to have many significant discontinuities (as should be obvious from the structure of the space of designs for software systems) and part of our task is to understand the nature of those discontinuities. the two spaces and their relationships. The variety of types of arrows is intended to show that there are different kinds and degrees of match between a region in design space and a region in niche space.

5 Trajectories in design space and niche space

One task for AI and related disciplines is to investigate possible trajectories in design space and in niche space, i.e. possible transformations from one design to another or from one niche to another.

This involves exploring and analysing possible forms of development, adaptation and learning within individuals and also possible types of evolutionary change.

Some changes occur within continuous regions of design space and niche space (e.g. smooth increases in speed of processing), while other trajectories cross discontinuities, e.g. introducing a notation or mechanism that (in principle) allows construction of nested symbolic structures of unbounded depth, going from a system of propositional logic to full predicate logic with quantifiers, or going from a purely reactive architecture to one that includes deliberative capabilities (described below).

There are some types of changes that can happen within a single individual, such as the changes from frog spawn to tadpole to adult frog, or the change from helpless human infant to naughty child, to sophisticated quantum physicist. Other types of trajectories in design space are not possible within an individual, but require evolution across gradually changing generations, or, in the case of artifacts, major re-engineering. For example, I suspect that there is no environmental manipulation that can transform a frog's egg into a giraffe. I do not know whether some sequence of evolutionary pressures could lead from a frog to a giraffe, possibly via regression to a simpler form (a common ancestor).

Whether any self-modifying artificial information processing system could start with the ability to write computer programs in assembly language and somehow extend itself by inventing languages like Algol, Simula67, Lisp, C++, Prolog, etc. or by inventing a new type of operating system for itself, remains an open research question, linked to other questions about mechanisms underlying human creativity.

Since all organisms form part of the environment for other organisms (including others of the same species) evolution in the design of one can constitute evolution in the niche for another, and *vice versa*. A study of which forms of co-evolution are and are not possible would be an essential part of the study of trajectories.

Another kind of trajectory is the evolution of a culture, i.e. the collection of concepts, knowledge, skills, norms, ideals, etc. shared (to varying degrees) among members of a community. There seem to be forms of learning that are possible in a culture but not in an individual (e.g. because they take too long to be achieved in one lifetime, or because they essentially involve interactions between individuals, such as social and political developments). Another way of thinking about this is to regard an enduring society as a particular form of self-modifying agent with a complex distributed architecture.

A different sort of question is whether a particular design permits instances to be assembled ready made in a laboratory or whether they would have to grow themselves. It may be physically impossible to assemble directly mechanisms that are capable of supporting certain kinds of functional architectures (e.g. assembling a fully functional adult human brain), because of the 3-D structural intricacies. This does not rule out the possibility of *growing* one in a laboratory, using sophisticated developmental and learning processes. But those are long term research issues, on which we can reserve judgement.

Whether a *software* equivalent to an adult human brain could be assembled in a fully functional form is another question. The answer may turn out to be "yes" in theory but "no" in practice, if the system is to be implemented in physical mechanisms and operate within human-like constraints of weight, physical size, speed of operation, and energy consumption. These are all questions on which opinions will differ until more research has been done.

6 Must designs be intelligible?

Another question on which there is disagreement is whether the provision of a large set of capabilities, such as those listed above, necessarily involves the creation of an *intelligible* design, with identifiable components performing separate tasks, or whether the functionality could sometimes (or always?) emerge only in a very complex and incomprehensible fashion from myriad interacting components.

For example, experimenters using genetic algorithms to evolve neural nets to control a robot sometimes create networks that work, but which seem to be impossible to understand (not unlike some legacy software which has grown over many years of undisciplined development).

This is related to the question whether a niche (i.e. a set of requirements) will always decompose into a collection of distinct capabilities which can be served by distinct components of a design, or whether there is always so much intricate "cross-talk" between requirements and between elements of designs that clean, intelligible, modular solutions will turn out to be impossible, except in relatively trivial cases.[1]

Even if designs are unintelligible at one level of description, there may be higher level descriptions of important features which can be discovered if only we develop the right sets of concepts. Cohen and Stewart [3] suggest that this emergence of higher level order is a feature of all complex systems, including biological systems.

7 How can an architecture be evaluated?

Evaluation of an architecture (or a generic design for a family of related architectures) can take different forms, depending on one's interests.

For instance, someone with a practical objective would be primarily interested in observable performance. This could include multiple dimensions of evaluation, involving input-output mappings, speed, running costs, generality, precision, accuracy, adaptability.

A much discussed (and maligned) criterion is the Turing test. The main point to note about this is that it corresponds to a tiny subset of niche space (even if interesting regions of design space are potentially relevant, as Turing claimed, at least implicitly). For someone interested in designs that fit other regions of niche space, the Turing test would be of limited value: a machine that passed the Turing test with flying colours might not be able to learn to fly an airliner safely, or to interpret the sensory information and control the movements of a robot.

Arguing about which performance criterion is correct is just silly: different criteria will be relevant to different scientific and engineering goals.

The task of designing a system satisfying observable performance criteria may lead to a concern with *internal* processes. For instance, whether a system can modify its performance by changing its strategies when things go wrong will depend on what sorts of internal monitoring, analysis and evaluation are possible, and what sorts of short term and long term internal self-modification are possible. This in turn will depend on the forms of representation and inference available, and the generative power of the internal building blocks.

Someone with a biological or psychological orientation, rather than practical engineering objectives, will have different criteria for evaluating models, for instance requiring a fairly close correspondence with information-processing states, and possibly even neural mechanisms, within the organism

[1] I've argued against certain sorts of modularity in vision, in [8].

being modelled. Detecting such a correspondence, or lack of it, may be very difficult, especially when the objective is to achieve a correspondence at a high level of abstraction compatible with significant differences in physical construction and differences in observable behaviour (just as different human beings sharing many design features will differ in their behaviour and capabilities).
A more general and ambitious scientific concern would be not just the evaluation of any particular model, or the study of any particular type of organism, but rather the comparative study of different architectures and their relationships to different niches. This could also include an interest in possibilities for change: i.e. a study of possible trajectories in design-space and niche-space, as described above. In particular questions about the *power* of an architecture may need to distinguish the power of the system at any particular time and the potential for increased power through learning and self-modification: consider the difference between a newborn human infant and other newborn mammals which walk, find the mother's nipple, and even run with the herd shortly after birth.

8 Designs for a new philosophy

This comparative analysis of types of designs and niches and their relationships is very close to old philosophical problems about the nature of mind, intentionality, consciousness, etc.
One difference is that whereas older philosophers used to ask questions like: "What is a mind?" or "What are the necessary and/or sufficient conditions for something to be conscious?" we can now ask "How many different kinds of minds are there and how do they differ in their architectures and their capabilities?" These questions unify philosophy, psychology, biology and AI. (Though we must resist any temptation to assume that the concept of a mind is initially clear, or that there are sharp boundaries between things with and things without minds!)
In philosophy, there is a long tradition of linking the possession of mental states (beliefs, desires, intentions, etc.) with rationality, and this tradition has recently manifested itself in Dennett's notion of the "intentional stance" and Newell's "Knowledge level" both of which require that actions be explainable in terms of beliefs and desires as if the agent were rational. However from our broader standpoint we can explore a variety of more or less "rational" architectures and assess them from different standpoints. E.g. for genes to perpetuate themselves it may be *essential* that agents sometimes behave in a manner that is not rational from the agent's viewpoint. There are many ways in which exploring design space can shed light on philosophical problems.

9 Is the task too hard?

Given the enormous diversity in both design space and niche space and our limited understanding of both, one reaction is extreme pessimism regarding our ability to gain significant insights. My own attitude is cautious optimism: let us approach the study from many different directions and with many different methodologies and see what we can learn. Even the discovery that a particular approach does not get very far is an advance in knowledge.
In particular, the Cognition and Affect group at Birmingham has been trying to use a combination of philosophical analysis, critical reflection on shared common sense knowledge about human capabilities, analysis of strengths and especially weaknesses in current AI systems, and where appropriate hints from biology, psychology, psychiatry and brain science, to guide a combination of speculation and exploratory implementation (e.g. using the general-purpose Sim_agent toolkit

[16]). The implementations inevitably lag far behind the speculation! The rest of this paper illustrates some of the speculation regarding functional decomposition [2]. I have speculated elsewhere about the diversity of forms of representation required in systems with human-like intelligence [3].

10 "Broad" agent designs

For now, let us ignore most of the types and levels of architecture and focus mainly on the highest level functional architecture: the global organisation of a collection of coexisting, interacting, capabilities, each of which may be described at a high level of abstraction, for instance, receiving or collecting information from the environment, analysing such information, interpreting the information; making plans to modify the environment, modifying the environment, monitoring modifications; generating new motivators, assessing motivators, working out costs and benefits of motivators, assessing likelihood of success, deciding whether to accept or reject them; monitoring internal processes, evaluating internal processes, modifying internal processes; and many more, concerned with different time-scales, different spheres of influence, different purposes. (Not all purposes need ultimately be those of the agent: e.g. much of animal behaviour serves the needs of a community, or a gene-pool, rather than the individual.)

This focus on the problem of combining a large number of diverse kinds of functionality, each of which may not (at first) be specified or modelled in much depth, has been dubbed the "broad and shallow" approach by the OZ group at Carnegie Mellon University [1].

11 Three levels of control

Within this framework I'd like to offer some speculations about the gross features of the human information processing architecture. These speculations are prompted by reflection on (a) many facts about human capabilities, (b) considerations regarding evolution of intelligence and (c) engineering design considerations inspired by reflection on limitations of current AI systems.

A brain is, above all, an information processing control system. I'd like to suggest that there are three rather different sorts of control, which might have evolved at different times.

11.1 1. A reactive subsystem

The first sort has been the focus of a lot of interest in recent years, in connection with "reactive" agents. In a purely reactive agent (or one sort of reactive agent) information is acquired through external sensors and internal monitors and propagates through and around the system, and out to effectors of various kinds, as indicated roughly in Figure 2.

This leaves open the possibility of some effects being counterbalanced by opposing tendencies, or some of the outputs of sub-components being gated or inhibited by others. Many different relatively unintelligent mechanisms of conflict resolution can fit into a reactive system. What a purely reactive system cannot do is explicitly construct representations of alternative possible actions, evaluate them and choose between them, all in advance of performing them.

[2] Reported in several previous papers [15, 7, 8, 9, 10, 2, 14, 17]. Compare [5].

[3] E.g. see [6, 11, 13]

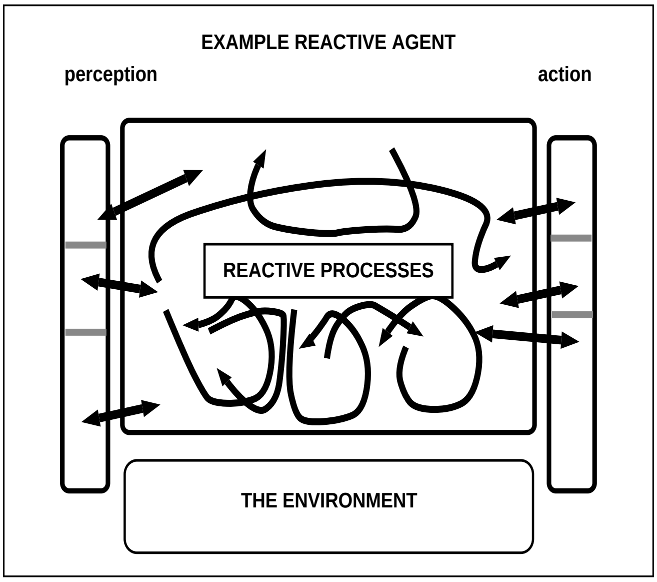

Figure 2: **An architecture for a reactive agent**

Processes occur in parallel in a reactive system because there are dedicated coexisting circuits. I presume there are many organisms like that (e.g. insects), and older, more primitive parts of the human brain are also like that.

In human beings, and possibly other animals, there are forms of learning, or rather training, that extend the capabilities of the reactive sub-mechanism. Thus we can distinguish designs for reactive systems that are largely static (apart from dynamic tuning of feedback loops perhaps), and designs that are extendable, possibly under the control of other mechanisms within the global architecture.

11.2 2. A deliberative subsystem

One of the major characteristics of a reactive system as conceived here is that all responses, whether internal or external, happen as soon as their triggering conditions are satisfied (provided that the response is not inhibited as a result of another reactive mechanism.) This principle of automatic triggering is independent of how the system is implemented, e.g. whether it uses a collection of neural networks, or condition-action rules in a symbolic rule interpreter, or something like procedure calls in a programming language, or just a hard-wired circuit.

If such a system is well matched to its niche, the fact that it is relatively inflexible and unintelligent is of no concern. It could be that insects are like this. Perhaps those mammals (e.g. deer) which are born with sophisticated capabilities that enable them to run with the herd also have an essentially reactive control system.

Such a system can break down when the pre-designed collections of conditions for triggering responses are confronted with new situations for which no appropriate responses are available. This is typical of the sort of niche that requires our second main type of control architecture, a "deliberative" architecture which is able to assemble new combinations of actions to cope with novel contexts, as indicated roughly in Figure 3.

In general the space of such combinations is explosive in its complexity[4], and that means that if

[4] If K choices have to be made from N types of components there will be of the order of N^K possible combinations.

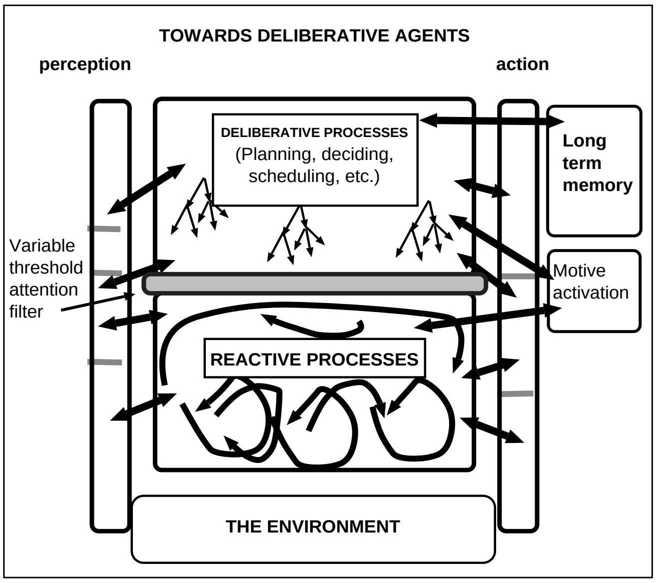

Figure 3: **A hybrid reactive and deliberative agent**

the new combinations have to be tried out by acting on them a very large number of experiments will be required, which may be both time consuming and very dangerous. So it is beneficial if the search can be done hypothetically, using some kind of model which is evaluated internally.

That sort of niche requires designs that include a type of memory in which temporary structures can be created, evaluated and then tried out. It may require storage of a number of different temporary structures, e.g. alternative plans that have to be compared in some way prior to selection. (This is the core difference between a deliberative and a purely reactive system.)

The processes which create, modify, compare, evaluate, select such new structures may themselves be implemented using more primitive reactive systems, which unlike the previous ones are primarily concerned with operations on an internal world rather than operations on the environment, though the result of their manipulations can be improved ability to operate on the environment.

This kind of deliberative mechanism, by definition, does not have pre-allocated resources for various functional capabilities: rather it is using a general subsystem to create and evaluate new capabilities including some which are then rejected.

There are many implications of this. In particular, because the same facility is being re-used for different sub-tasks, questions about resource limitations arise, which are not relevant to reactive systems where dedicated circuits exist for the different sub-capabilities. Other obvious questions arise, such as whether and how these newly created structures can be stored and retrieved in similar contexts in future.

Yet another problem is whether the re-activation of a previously constructed plan necessarily makes use of the same mechanisms as create new solutions to problems, so that it is not possible then to use the deliberative mechanism to solve a new problem while one of its previous products is being used.

A possible solution is to transfer newly constructed solutions to the reactive subsystem, where they can in future be run in parallel with new deliberative processes. This seems to be a feature of many kinds of human learning, including familiar examples such as learning to drive a car, learning to read text or sight read music, becoming a fluent programmer, learning many sporting skills.

The diagrams above and below are intended to indicate that perceptual subsystems and action

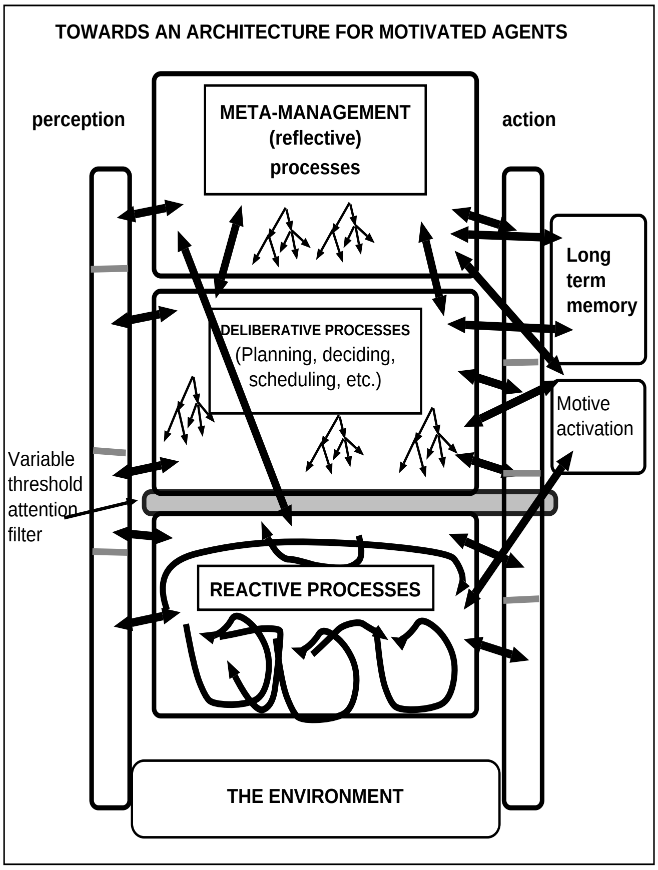

Figure 4: **Adding a meta-management layer**

subsystems can develop different levels of abstraction corresponding to the different requirements of the more central architectures to which they are connected. For instance a deliberative architecture may require the perception of abstract "affordances" in the environment in order to perceive the possibility of a particular step in a plan that is being considered. Moreover social agents and agents that interact with other animals may need to be able to infer complex internal mental states from observed facial expressions posture, gestures, etc. Our ability to see a face as happy or sad or threatening would be examples.

In previous papers my colleagues and I (largely inspired by [5]) have been exploring some of the consequences of the division of labour between a reactive system and a deliberative system, including the implications of concurrent triggering of new motives by the reactive system, sometimes when the deliberative system is overloaded, necessitating some sort of "attention filter" to protect processes that are urgent, important and difficult. Some emotional states can be interpreted as arising out of "perturbances" in such an architecture [17].

11.3 3. A meta-management subsystem

The third sort of control system, which we have previously described as a meta-management system (e.g. [2, 14, 17]) is concerned with monitoring and control of the deliberative mechanism, as indicated in Figure 4.

The idea is that just as a reactive system may suffer from excessive rigidity in a changing environment, so may a deliberative mechanism. In particular since the environment of the deliberative system is in part the internal architecture of the agent, and since that environment changes as the products of the deliberative system are stored and made available for future use, it is very likely that what works in the early stages of an agent's development may not be very good at much later stages. For this and other reasons it would be useful for internal monitoring mechanisms to be able to keep records of processes, problems, decisions taken by the deliberative mechanism, and perform some kind of evaluation, relative to high level long term generic objectives of the agent (some of which might be determined genetically, and some of which might be learnt in some way, including possibly being absorbed from a culture).[5]

Generic objectives could include such things as not failing in too many tasks, not allowing the achievement of one goal to interfere with other goals, not wasting a lot of time on problems that turn out not to be solvable, not using a slow and resource-consuming strategy if it turns out that a faster or more elegant method is available, and detecting possibilities for structure sharing among actions.

Although such a meta-management system may have a lot in common with a deliberative subsystem, the point of making the distinction is that the deliberative mechanisms *could* exist without the kinds of self-monitoring and self-assessing capabilities just described. In fact, I conjecture that comparative studies will show that that is the case in many animals. Moreover just as deliberative mechanisms can vary in their scope and sophistication so also can meta-management mechanisms.

It might be argued that if meta-management is needed then so also is meta-meta-management, and so on. However, the three kinds of subsystems may suffice if the kinds of self-monitoring and self-modifying capabilities which I've ascribed to the third layer can be applied to itself. We then need no new kind of subsystem.

There are many unanswered questions. For example, experience with computing systems suggests that it is difficult or impossible for everything to be monitored: in fact in the limiting case that would produce an infinite regress of monitoring mechanisms. It may also be the case that there are incompatibilities between the requirement for certain processes to be internally monitored and the requirement for them to run fast on dedicated circuits. This could imply, for example, that the self-monitoring mechanisms used for meta-management cannot have direct access to all the details of the workings of the reactive system.

To overcome this, special additional circuits within the reactive system might be used to transfer information about low level processes to deliberative and meta-management processes which can use it for high level evaluations of current activities. Such "internal perception" mechanisms could simplify and abstract, if that suffices for the job, in which case higher levels will have access only to incomplete and possibly misleading information about what is going on, not unlike senior management in a large organisation!

These design problems are relevant to a lot of contemporary discussions about consciousness, qualia, and the role of introspection. My own view is that the vast majority of what is written on such topics (even by distinguished scientists) is of dubious value because it has not been based on an implementable theory of the architecture which could support the concepts used by the discussants. (I am not restricting consideration only to computational implementations.)

---

[5] For more on reasons for self-monitoring see [4].

12 Further questions

The sort of discussion presented here needs to be combined with the more familiar AI research on formalisms and algorithms. It could well turn out that quite different formalisms are suited to the different tasks. Different formalisms and ways of manipulating them may require the existence of different kinds of representational media.

In particular a reactive subsystem may be able to use forms of representation and control which are not suited to a deliberative system, including, in the extreme case, hard-wired circuits and reflexes. If so that raises interesting problems about what happens when as a result of training new structures created by the deliberative system get implanted (or transplanted?) to the reactive subsystem.

Is the very old idea that some forms of learning are a bit like compiling from a high level to a low level language supported by this?

Alternatively might it be that the very information structure that is created by a deliberative mechanism can also be used by a reactive system, but in a far less flexible (though speedy) fashion?

Too often it seems that debates about mechanisms and formalisms (e.g. logical notations vs neural nets) are conducted in a spirit in which issues of partisanship, or fashion, have more influence than scientific considerations. I suspect that by asking how all the various components can be put together into complete working systems we may be able to make more progress with such problems and even learn that instead of having to choose between apparently incompatible options we have to use both, but in different parts of the system. In short, debates about which sorts of formalisms are best should be replaced by investigations mapping formalisms to tasks, within the more general study of relations between designs and niches.

13 Other aspects of the architecture

Claiming that an architecture has reactive, deliberative and meta-management sub-systems does not imply that each of these is a monolithic mechanism, or that everything in the architecture must fit neatly into one of these categories.

Perception is an interesting example. In an agent whose complete architecture is reactive, perceptual mechanisms will use fixed algorithms for analysing their input and determining what should be sent on to other parts of the system. Where the architecture includes a deliberative component, however, a perceptual system could have a dual role, namely both feeding information directly into the reactive subsystem and also collaborating with the deliberative system when it constructs and evaluates alternative possible action plans. A chess-player working out what move to make will often find it useful to stare at the board and use it as an extension of short term memory (though a more advanced player can do this all internally). Similarly an animal considering how to pick something up, or which route to take across a cluttered environment, may find that the problem is easier to solve while the environment is visible, again because the perceptual structures form part of the re-usable short term memory structure required for creating and evaluating options.

The often rediscovered fact that humans use spatial representations for solving many kinds of problems, including some very abstract problems, may be a manifestation of the overlap between a spatial perception mechanism and the deliberative mechanism. On the other hand, the visual feedback that allows smooth and rapid movement of a hand to pick up a cup could be an example of a deep connection between spatial perception and some reactive mechanisms.

If all this is correct, perceptual mechanisms are neither entirely in the reactive subsystem nor entirely in the deliberative subsystem. Similar comments could apply to the motor output system, if the reactive subsystem sometimes controls it and at other times the deliberative subsystem takes over, or if both can be simultaneously involved in different aspects of the control of behaviour, e.g. thinking about phrasing and dynamics by performing a well-rehearsed piece of music.

A different sort of point concerns the question whether within the perceptual system there is a need for a distinction between reactive and deliberative subsystems. It may be that the perception of complex structures (e.g. hearing grammatical sentence structures, or seeing a complex piece of machinery) requires some ambiguities of parsing or local interpretation to be resolved by temporary construction of alternatives which are compared. If so, a perceptual mechanism may need to include something analogous to deliberative mechanisms, though possibly tailored specifically to the tasks and forms of representation in that mode of perception. (This was taken for granted in much AI vision research in the 1960s and 1970s, but later went out of fashion.)

14 Motivation

I have hinted that new motives can be generated asynchronously in different parts of the system. How all these motives are managed is a complex topic that has not been investigated much in AI[6]. In psychology and neuroscience, I have the impression that much of the study of motivation, emotions and related states and processes, has assumed that humans are essentially the same as other animals, such as rats. This assumption may be misleading. Motivational processes in an agent whose deliberative mechanisms can explicitly represent the long term future may have significant additional complexity compared with the processes that occur in a rat, for example. Can the latter feel humiliated, guilty, awe-struck or driven by a long term ambition?

Agents that can learn through positive and negative reinforcement will have their motivational mechanisms linked to their learning mechanisms so that rewards and punishment bring about changes. Agents that also include meta-management, i.e. agents that are capable of monitoring, evaluating, and modifying high level aspects of their own internal processes, will be capable of having very abstract types of motivation that simply could not occur in agents with simpler architectures, for instance the desire to be an honest and generous person.

The three layers support very different sorts of mental processes, including motivational and emotional processes, a fact that has not been noticed among emotion theorists, leading to a plethora of different definitions of "emotion" and unrelated theories of emotion, which appear to contradict one another, but are actually talking about different things. Our diagnosis is that those who stress emotions based on the limbic system and observable in rats and most other animals are studying effects of the reactive layer. Those who stress emotions such as apprehension, disappointment and relief, related to phases in the execution of plans, are studying effects of the deliberative layer. By contrast poets, novelists and those who study emotions involving loss of control of thought processes (e.g. our work on grief and perturbances [17]), are studying processes involving the reflective, or meta-management layer.

There is much more to be said about motivation, moods, character, personality, and the like. In particular, requirements for concurrency and independence of various subsystems can lead to a variety of kinds of states in which subsystems disturb one another, possibly producing less than optimal global performance. Some human emotional states, including states that are too sophisticated to occur in rats, may be like that.

[6] Though see [2] and references therein.

Some AI researchers believe that it should be the goal of AI to design agents that overcome human limitations while displaying all their strengths. This may not be possible if some of the limitations are inevitable consequences of the mechanisms and architectures required to produce those strengths.

15 Conclusion

I have tried to outline a methodology which takes account of the existence of niche space and design space and their relationships.

I have also tried to illustrate the application of this methodology to the analysis of a particular class of designs and niches, showing how this might be achieved using an architecture which (among other things) has reactive, deliberative and meta-management components (a trio that may correspond loosely to old and familiar concepts from philosophy, psychology and common sense).

What I have not done is to spell out examples of complete working architectures to show what kinds of ontologies for mental states and processes they support and how well they can explain sophisticated aspects of human mentality. This is ongoing work.

16 Acknowledgements

This research was funded in part by: the Renaissance Trust, the UK Joint Council initiative on Cognitive Science and HCI, and DRA Malvern. I have received much help from colleagues and students at Birmingham, including Luc Beaudoin, Chris Complin, Darryl Davis, Glyn Humphreys, Brian Logan, Riccardo Poli, Christian Paterson, Ed Shing, Tim Read and Ian Wright.

This is a slightly expanded version of an invited paper presented in August 1996 at the Cognitive Modeling Workshop at AAAI96, Portland Orego.

References

[1] J. Bates, A. B. Loyall, and W. S. Reilly. Broad agents. In *Paper presented at AAAI spring symposium on integrated intelligent architectures*, 1991. (Available in SIGART BULLETIN, 2(4), Aug. 1991, pp. 38–40).

[2] L.P. Beaudoin. *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham, 1994.

[3] J. Cohen and I. Stewart. *The collapse of chaos*. Penguin Books, New York, 1994.

[4] J. McCarthy. Making robots conscious of their mental states. In *AAAI Spring Symposium on Representing Mental States and Mechanisms*, 1995. Accessible via http://www-formal.stanford.edu/jmc/consciousness.html.

[5] H. A. Simon. Motivational and emotional controls of cognition, 1967. Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979.

[6] A. Sloman. Interactions between philosophy and ai: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, London, 1971. Reprinted in *Artificial Intelligence*, pp 209-225, 1971, and in J.M. Nicholas, ed. *Images, Perception, and Knowledge*. Dordrecht-Holland: Reidel. 1977.

[7] A. Sloman. Motives mechanisms and emotions'. *Emotion and Cognition*, 1(3):217–234, 1987. Reprinted in M.A. Boden (ed), *The Philosophy of Artificial Intelligence*, 'Oxford Readings in Philosophy' Series, Oxford University Press, 231–247, 1990.

[8] A. Sloman. On designing a visual system (towards a gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337, 1989.

[9] A. Sloman. Prolegomena to a theory of communication and affect. In A. Ortony, J. Slack, and O. Stock, editors, *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, pages 229–260. Springer, Heidelberg, Germany, 1992.

[10] A. Sloman. Prospects for ai as the general science of intelligence. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, and A. Ramsay, editors, *Prospects for Artificial Intelligence*, pages 1–10. IOS Press, Amsterdam, 1993.

[11] A. Sloman. Semantics in an intelligent control system. *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering*, 349(1689):43–58, 1994.

[12] A. Sloman. Exploring design space and niche space. In *Proceedings 5th Scandinavian Conference on AI, Trondheim*, Amsterdam, 1995. IOS Press.

[13] A. Sloman. Musings on the roles of logical and non-logical representations in intelligence. In Janice Glasgow, Hari Narayanan, and Chandrasekaran, editors, *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, pages 7–33. MIT Press, 1995.

[14] A. Sloman. What sort of control system is able to have a personality, 1995. Available at URL ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/ Aaron.Sloman.vienna.ps.Z, (Presented at Workshop on Designing personalities for synthetic actors, Vienna, June 1995).

[15] A. Sloman and M. Croucher. Why robots will have emotions. In *Proc 7th Int. Joint Conf. on AI*, Vancouver, 1981.

[16] A. Sloman and R. Poli. Sim_agent: A toolkit for exploring agent designs. In Mike Wooldridge, Joerg Mueller, and Milind Tambe, editors, *Intelligent Agents Vol II (ATAL-95)*, pages 392–407. Springer-Verlag, 1996.

[17] I.P. Wright, A. Sloman, and L.P. Beaudoin. Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126, 1996.

For more on our work see the project ftp directory:
ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect