To appear in:
Proceedings Workshop on Designing personalities for synthetic actors
Vienna, June 1995
Ed. Robert Trappl

WHAT SORT OF CONTROL SYSTEM IS ABLE TO HAVE A PERSONALITY?

Aaron Sloman
School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, England
A.Sloman@cs.bham.uk.ac
Phone: +44-121-414-4775

Abstract

This paper outlines a design-based methodology for the study of mind as a part of the broad discipline of Artificial Intelligence. Within that framework some architectural requirements for human-like minds are discussed, and some preliminary suggestions made regarding mechanisms underlying motivation, emotions, and personality. A brief description is given of the 'Nursemaid' or 'Minder' scenario being used at the University of Birmingham as a framework for research on these problems. It may be possible later to combine some of these ideas with work on synthetic agents inhabiting virtual reality environments.

1 Introduction: Personality belongs to a whole agent

Most work in AI addresses only cognitive aspects of the design of intelligent agents, e.g. vision and other forms of perception, planning, problem solving, the learning of concepts and generalisations, natural language processing, motor control etc. Only a tiny subset of AI research has been concerned with motivation and emotions, or other things that one might regard as relevant to personality.

Partly inspired by Simon's seminal 1967 paper, I have been trying since the late 70s, in collaboration with various colleagues and students, to address these issues. They are intimately bound up with deep and difficult problems about how human minds work, and I don't expect answers to be found in my lifetime, though that's no reason for not trying to make a beginning. Doing this requires thinking about the design of 'complete' agents (whether human-like or not), not just specific cognitive mechanisms. That's a very difficult task, since we know so little about so many of the components and underlying implementation mechanisms. Nevertheless, by combining design of 'broad' but temporarily 'shallow' architectures with various other kinds of research on more detailed mechanisms we can hope gradually to make progress towards complete and realistic designs and theories.

In order to make clear the framework within which I am working, I'll start by making some general comments on the scope and methodology of AI. I'll then describe a scenario within which some of us are exploring possible architectures to account for aspects of motivation, emotion and personality. And I'll then sketch some preliminary ideas about an explanatory architecture, which is not offered as a complete specification, but a partial, high level overview of a family of possible architectures. At the end I have included an edited transcript of some tape recordings following the discussion after my presentation at the workshop in Vienna in June 1995, as this may help to remove some common misunderstandings.

Why include a section on the goals of AI? Part of the reason for this is that most people think of AI in terms that are too narrow. I shall try to offer an alternative vision of AI that is broad and deep, within which a study of personality can be accommodated. It also helps to identify many unsolved problems and weaknesses in current research.

2 How should we identify aims of AI?

There are various approaches to defining the aims of AI, including the following:

1. Try to articulate what you yourself think you are doing and what larger set of goals it fits into. This is what many AI practitioners do. Some are unaware of what lots of others do.

2. Repeat some definition of AI that you have read, especially if originally produced by one of the founders or 'gurus' of AI. This is what many undergraduates, recent graduates, recent recruits, journalists, and outsiders do.

3. Look at what actually goes on in AI conferences, AI journals, books that claim to be on AI, AI research labs: Then try to characterise the superset. This could be what a sociologist or historian of science might do. Many AI people now tend to go only to their specialist conferences and read only their specialist journals, so they lose the general vision. So also do many external commentators on AI.

4. Like (3) but instead of simply characterising the superset, try to find some underlying theme or collection of ideas which could generate that superset.

The last is what I have tried to do. This is what I have learnt:

2.1 What are the aims of AI?

AI is (as I have claimed in various previous publications):

The general study of self modifying information-driven control systems,

- both *natural* (biological) and *artificial*,
- both *actual* (evolved or manufactured) and *possible* (including what might have evolved but did not, or might be made, even if it never is.)

I include the study not only of individual agents, but also societies and the like: social systems add new constraints and new design possibilities, relating to communication, cooperation and conflict. By the 'general study', I mean to include: not just the creation of any particular such system, but an understanding of what the options are, and how they differ and why, including what the trade-offs are.

From this standpoint, Cognitive Science is the subset of AI that studies human and other animal systems. AI thus defined has a number of sub-activities, not often thought about, which I'll now summarise.

3 Sub-tasks for AI

This general study of behaving systems is *not* a search for a particular algorithm or design. Even if we had a truly marvellous AI system equalling or surpassing humans in many respects, that would not be enough. For we'd need to be able to *understand* what we had done, and why certain aspects of the design were good and why alternatives would not be as good. Such understanding involves knowing not only what a particular design is and what it can do, but also how it relates to other possible designs. It also involves knowing which aspects of the implementation are essential and which are not.

In short we need to understand the space of different possible architectures, i.e. the different regions of design space and their properties. In particular, for different classes of designs and mechanisms we need to understand what are they good for or not so good for. Namely, which collections of requirements do different designs fit into? These questions can arise at many different design levels. (Some of the issues are discussed, though in the more general context of understanding complexity and simplicity, by Cohen & Stewart 1994.)

3.1 What is a niche?

Using terminology from from biology, and slightly generalizing it, I use the word 'niche' to refer to a collection of requirements for a working system, such as an engineering requirements specification. Any particular design may fit a niche more or less well.

A niche is not just a physical thing nor a geographical region. A chimpanzee, a squirrel, a parrot, or a flea might each be placed on the same branch of the same tree in the same forest, yet each would be in a different niche from the others. E.g. they need to perceive different things in the environment, and when they perceive the same thing they may use the information for different purposes. Similarly, different software packages, such as a compiler, an editor, a database, a spreadsheet, an internet browser, will all occupy different niches, within the same computer.

A niche is an abstraction, not a portion of the world. A particular physical location may instantiate several different niches at the same time. The bee and the flower which it pollinates, are located in different niches, though their niches are related: anything meeting the functional requirements of each of them will help to define part of the niche for the other.

The specification of the niche for a particular type of agent could include: (a) the ontology of the environment, as sensed and acted on by the agent, (b) the dynamics possible within that ontology (which events and processes can occur), (c) the means of sensing the environment available to the agent, (d) the types of actions required of the agent, and (e) a collection of possible tasks and constraints, where the set of actual tasks may be dynamically changing. Exactly what is included will depend on how precisely the niche is specified.

For example the tasks might include finding and consuming enough food to keep alive, finding a mate, reproducing, looking after young, learning about a social system, fitting into the social system, etc. Some constraints are imposed by laws of nature, e.g. physical constraints, and others by a legal system ruling out certain means of achieving goals, or a set of conventions for communication. A niche may be made more constraining by adding additional requirements, e.g. specifying what sort of food is to be used, or which other sorts of agents are to be aided in some way. Because any particular portion of the world can instantiate several different sets of descriptions simultaneously, it can instantiate several niches simultaneously.

Some niches are determined more or less explicitly in advance by human engineers (or their customers) and guide their design work. Other niches are implicit in a collection of evolutionary pressures that operate on a class of organisms. Just as humans can design things with complementary roles (e.g. plugs and sockets, compilers and machines) so naturally occurring implicit niches may complement one another. One way of looking at an ecology is as a collection of interacting niches, which may be changing dynamically. Different aspects can change independently, e.g. changing climate alters the requirements regarding discovery or creation of suitable nests or shelters, whereas a changing culture, or changing amounts and kinds of knowledge alter the requirements for individual learning, and collaboration or competition with others in the society. In a complex society with systematic division of labour, different social roles will require different individual types of motivation, preferences, ambitions, likes, dislikes, etc. I.e. different personalities will be required. An extreme example is the difference in reproductive roles.

3.2 What is a design?

A design, like a niche, is not something concrete or physical, though it may have physical instances. A design is an abstraction which determines a class of possible instances, and where a design is capable of being instantiated in different ways, there will be more specific designs corresponding to narrower sub-classes of possible instances.

Talk about 'designs' does not imply anything about the *process* of discovery or creation of designs. ('Design' can be a noun as well as a verb.) Design production does not have to be top-down: it can be bottom-up, middle-out, or multi-directional. Arguing that only one approach will work, as some defenders of genetic algorithms or neural nets do, is silly: all approaches are liable to 'get lost' searching in design space.

There is no one true road to understanding: we have to follow several in parallel and share what we learn. The approach can be empirical or theoretical. When it is theoretical it may be either intuitive and vague or formal and precise, making use of logic and mathematics. It may but need not include the creation and study of instances of the design. When instances are created (i.e. a design is implemented in a working system), this is often part of the process by which we understand the problem, rather than our top level goal. Investigation by implementation is very common in AI, and is partly analogous to the role of thought experiments in physics.

Designs include specifications of architectures, mechanisms, formalisms, algorithms, virtual machines etc. Where a design specifies a complex structure with interacting components, it will need to include not only structural features but also the behavioural capabilities of the components and their possible forms of interaction, i.e. their causal powers or functional roles within the whole system.

Differences between designs include both (a) different ways of refining a common more general design, e.g. starting with a general parsing mechanism and then applying it to two specific grammars, to produce parsers for those grammars, and also (b) differences that are due to different implementations in lower level mechanisms, such as using different programming languages, or different compilers for the same program, or compiling to different machine architectures, or implementing the same machine architecture using different physical technologies.

In many cases a particular design D can be implemented in different lower level mechanisms. In that case we say D is a design for a *virtual* machine, and different instances of that virtual machine may occur in quite different physical systems, for instance when different physical technologies are used to implement the same computer architecture, e.g. a VAX or a SPARC architecture. Insofar as the process of biological evolution can be seen as a mechanism that explores design space it seems to make use of very different levels of implementation, most of them being the result of previous designs, whether of reproductive mechanisms, chemical structures and processes, neural mechanisms, or mechanical structures.

Often it is impossible or difficult to create instances of a new design directly, so part of what has to be designed includes new production processes. This point is often stressed (e.g. by Cohen & Stewart 1994) in criticising the notion that DNA fully determines the development of an embryo, for that ignores the role played by the mechanisms that 'interpret' the DNA.

Human design capabilities are enhanced by development of new design and manufacturing tools. Thus closely associated with any particular design may be a set of more generic 'meta-designs' for design and production mechanisms. The latter have a niche that is determined by the kinds of designs they are required to enable or facilitate.

The less specific a design the more scope there is for varying the implementation details. One of the things we don't yet understand well is which classes of designs are neutral between very different kinds of implementations, e.g. which high level designs for intelligent human-like agents are neutral as to whether the components are implemented in a collection of neural networks and chemical soups or in a collection of symbol manipulating systems, or some mixture of mechanisms.

So, we don't yet know which high level aspects of the design of a human mind are neutral between implementation on human brains and computer-based implementation, though much prejudice one way or the other abounds. Answering that question is among the long term objectives of AI as defined above.

A related question is the extent to which complex behavioural capabilities can be explicitly built in in advance, or whether mechanisms capable of implementing such capabilities cannot be directly programmed, but must 'program' themselves through processes of development, learning and adaptation. E.g. it may be physically impossible, in any kind of factory, directly to assemble a fully formed adult human brain with all the information needed for normal adult functioning already in it.

In that case any system containing such a brain will have to have learnt a great deal for itself. Thus part of a requirement for its early personality will be a set of motivations and capabilities capable of driving such a learning process.

If we wish to understand how to give a synthetic agent a personality we need to understand what sort of niche makes having a personality relevant to the requirements the agent has to fit into, and what sorts of designs are capable of meeting such requirements. I've tried to show that that is a far more complex question than it might at first appear.

3.3 Studying 'niche-space' and 'design-space'

The general study, which I have claimed constitutes the scope of AI as it is actually practised in all its forms, involves at least the following aspects, though not all are found often in AI work.

1. The study of 'niche-space'

This is the study of collections of requirements and constraints for agent designs, each collection being a 'niche'. Besides particular niches we need to understand dimensions in which niches can vary, and also the dynamics of changes in interacting niche systems. Although this is not very often made explicit, there are examples in AI research, including the study of different requirements for

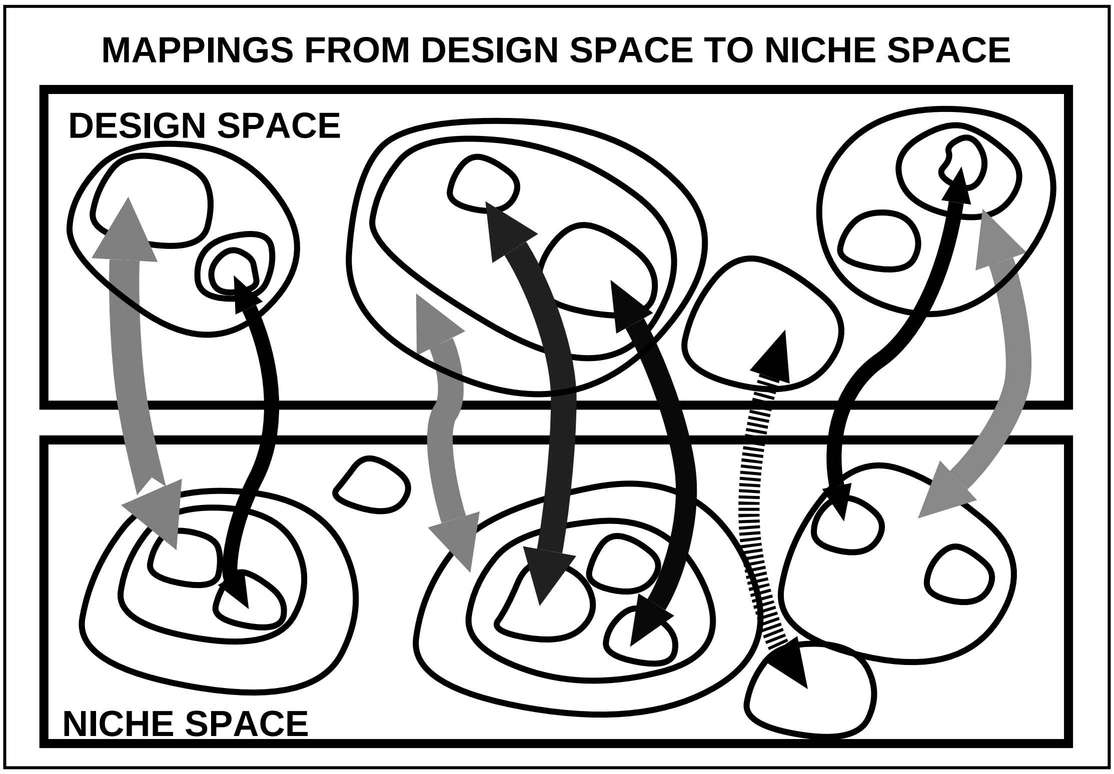

Figure 1: **Mappings between design space and niche space**
A niche is a set of requirements. Mappings can vary in degree and kind of goodness. Various kinds of trajectories in design space and niche space are possible.

learning systems, and the analysis of different sorts of perceptual tasks (e.g. Marr 1982, Sloman 1989). Marr misleadingly described this as the computational level of analysis. Many of the papers by myself and my colleagues are concerned with requirements for motivational mechanisms in human like systems (e.g. Sloman 1978 (chapter 6), Sloman and Croucher 1981, Sloman 1987, 1992, 1993, Beaudoin 1994).

2. The study of 'design-space'

This includes analysis and comparison of design possibilities at different levels of abstraction, including high level architectures, low level mechanisms, forms of representation and reasoning, types of long term and short term memory, types of perceptual mechanisms, types of learning mechanisms, and so on. This is found in much of the discussion in research papers comparing work by different authors. Unsurprisingly, it is rare to find discussions of designs or requirements for complete agents.

3. The study of mappings between design-space and niche-space.

These mappings are correspondences between designs (or classes of designs) and sets of requirements. For any given region of niche-space there are usually alternative designs, none fitting the niche perfectly. The different styles and widths of arrows in Figure 1 are meant to indicate different kinds and degrees of match.

In particular, there is no 1 to 1 mapping. There are many trade-offs and compromises, and no unique criterion of optimality for satisfying a niche. Neither is there a simple numerical measure of goodness of fit between a design and a niche. Design D1 may be better for niche N than design

D2 in some ways, and worse in others. E.g. it may be able to catch a wider range of edible prey because it can run faster, but not be so good at distinguishing edible from inedible items, because of its perceptual limitations. A full analysis of the variety of mappings requires an analysis of the logic of 'better', which is a far more complex concept than most people realise. E.g. we need to understand how to handle multiple partial orderings concerned with different types of comparison.

4. Analysis of different dimensions and levels of abstraction.

This involves looking at niches and designs from the standpoint of several different disciplines (e.g. neural, psychological, social, philosophical, linguistic, computational, etc.)

5. The study of possible trajectories in niche-space and design-space

AI, like many other sciences, is concerned with processes that are extended in time. When a complex system interacts with its environment that is external behaviour. In many cases this will involve internal behaviour (e.g. compression and changing stresses within a bouncing ball). Some internal behaviour is information processing, including processes of development or learning, such as happens in a foetus as it develops to a normal infant, or happens in an infant after birth. Many of these changes produce changes in the capabilities of the system. These are changes in design, even though no external agent has redesigned anything.

In other words, something that develops, or learns, follows a trajectory in design space. And because it can meet different sets of requirements and satisfy different constraints as a result of the changes, there are corresponding trajectories in niche space. (Though remember we are not talking about movement of a point in either space, but movement of some sort of region, which may have fuzzy boundaries.)

Explorations of learning systems then, are concerned with the study of mechanisms that can move themselves around design space (and niche space) partly as a result of interacting with an environment. The trajectories that are possible may depend on the particular environment - so from that point of view a niche may be part of the 'design' of a larger system.

One of the interesting questions to be investigated is which sorts of trajectories are possible and which are not. Just as some designs may have features that make direct implementation impossible, so that self-adaptation and learning during development are required, it could also turn out that there are some trajectories in design space that are simply impossible for any individual, though they can be achieved by evolutionary processes operating on a gene pool distributed in a collection of individuals.

For example, given the physics and chemistry of our universe, it seems to be impossible for for any individual to transform itself from an elephant into an ape, or from a flea into a human being, although such transformations might be logically possible. It's likely that neither organism includes the potential for such drastic changes, even if they are produced by gene pools which do.

There may also be types of transformations of information processing capabilities that are not possible for an individual. For example it may be impossible for a new born mouse to learn to understand any human language, even though a new born human can, and a mouse and a human share an enormous biological heritage.

Thus some trajectories within design space and niche space may be possible for a gene pool but not for any individual agent.

It might also turn out to be impossible for some of the forms of conceptual development exhibited in the history of human science and culture to occur within any one individual, no matter how long that individual lives. Perhaps a mixture of interleaved individual development and social learning is required to produce transitions such as those leading from ancient Greek science to quantum physics. For example this might be the case if the process necessarily involves 'bootstrapping' through mistaken or confused cognitive states which, once they have been entered cannot be left, even though they may be part of the environment in which a new generation learns while avoiding those states.

The same may turn out to be true for evolution of tastes, aesthetic preferences, art forms, moral values, and types of personalities. These are all topics for further study under the general heading of exploration and analysis of possible trajectories in design space and niche space.

Figure 1 gives a very rough indication of the sort of thing I have been discussing, though the use of a 2-D surface oversimplifies by suggesting that there is a unique level of analysis for designs and niches. This is not the case.

3.4 Discontinuities in design space and niche space

A point that is often not noticed, and which is important both for AI and for the theory of evolution is that the spaces are discontinuous.

Changes in designs can include either continuous variation or discontinuities. Some regions of design space have smoothly varying properties, e.g. changing speed, electrical resistance, or fuel consumption, or size. Others involve small or large discontinuities, like the change from having one wheel to having two, or three or four. In information processing systems, there are a lot of discontinuities in design. If you remove a conditional branch from a program, that's a discontinuity. You can't put back a half of it or a quarter, or an arbitrary fraction.

On the other hand, some discontinuous spaces may be capable of being embedded in continuous spaces, as the integers are embedded in the reals. Thus (as pointed out in (Sloman 1994b) a feed-forward neural net can be thought of as a large collection of condition action rules, all activated in parallel, with many conditions sharing certain actions, and many actions sharing certain conditions, where the weights determine a degree of influence of a particular rule. So in this case the degree of influence between a 'condition' and an 'action' can vary continuously between 0 and some significant value, even though such variation is not possible for conditional branches in ordinary software.

In conventional software, it may be possible to get something approximating such continuous variation by adding a randomiser to each condition, and gradually changing the probability with which the condition will be triggered. Nevertheless the difference between the design that allows a certain condition to play a role, however small, and the design that has no such condition at all, is a discontinuity. A structural change is needed to get from one to the other.

This is one example of a research topic regarding the structure of design space. We need to find out how many kinds of discontinuities there are, and which, if any, of them can be embedded in more complex designs that smooth out the discontinuities. (One reason for doing the latter is that it may allow 'hill climbing' in the search for good designs, e.g. using evolutionary algorithms, something I've learnt from my colleague Riccardo Poli.) Where there are irreducable discontinuities we need to understand how they relate to possible trajectories in niche space and to developmental or evolutionary mechanisms that are capable of producing such discontinuities. (It is not always acknowledged that Darwinian evolution *requires* discontinuous change between generations, even though the discontinuities may be small.)

3.5 AI and Philosophy

I hope the discussion of design and niche spaces and possible trajectories makes it clear why it is too limiting to conceive of AI as the search for any particular design, or a class of designs meeting any particular set of requirements. That may be a useful practical goal, but it is not enough for a deep study of mind.

In particular, the design of any particular architecture but a part of a broader study, which is to try to find out what sorts of architectures are possible, and how they meet different sets of requirements, i.e. how different areas of design space correspond to different areas of niche space.

This enables us to generalise the old philosophical question ‘What is a mind?’ and replace it with: ‘What sorts of minds are possible, and how are they different and what sorts of mechanisms and evolutionary or developmental or learning processes can bring them about?’

Whereas older philosophers tried to say ‘A mind has to be this, that or the other’, I suggest the answer has to be of the form ‘Well, if you design the thing this way, you have one kind of mind, if you design it that way, you have another kind of mind’ and so on.

It’s the differences and similarities between the different designs and how they relate to different niches that are important, not necessary or sufficient conditions.

In order to pursue this study we need a set of techniques and conceptual tools. I’ll now describe some of them.

4 Requirements for a study of design space and niche space

At the very least we need the following.

4.1 A language for describing niches (sets of requirements)

Some of the work by engineers in developing formalisms for expressing requirements may be helpful. Similarly some of the work done by biologists in comparing the niches of different but related organisms may be helpful. I suspect both have barely scratched the surface of what is required. Any satisfactory language will have to take account of the fact that a niche is an abstraction, not a physical environment.

Moreover, a niche has to be described from the ‘viewpoint’ of a type of agent.

A part of a design may correspond to part of a niche (as lighting system, fuel delivery system and steering mechanism in a car each comprises a part of the total design meeting different sub-requirements). Within each sub-niche and sub-design further decomposition is possible, e.g. the lighting system includes control switches, wiring, lamps, reflectors, power source, etc. Some sub-systems may share components with others, e.g. the battery is a part of several different sub-systems, and therefore occupies several niches simultaneously.

Some aspects of a sub-niche may be defined by ‘other’ features of an agent – the same agent or other agents. For example, an individual that cannot think very quickly but lives in an environment in which things change rapidly may need the ability to solve some problems very rapidly without deep thinking. (A standard answer is pattern recognition capabilities.).

The language for niches will have to evolve as our theories of possible designs evolves.

4.2 A language for formulating designs

Much of the important specification of designs for behaving systems is concerned with what I have called (Sloman 1994b) the 'information level'. This is

(a) Below Newell's 'knowledge level and' Dennett's 'intentional stance' level of description

(b) A level concerned with designs (i.e. part of Dennett's 'design stance') but also involving semantic content of information that is acquired, created, manipulated, stored, or used.

(c) A level at which information can be processed without presupposing rationality (as Newell's 'knowledge level' and Dennett's 'intentional stance' both do. In particular, evolution or other designers can produce systems that work in given situations even though they are not rational, e.g. because they blindly follow rules. In particular, where an agent is part of a larger system, e.g. a society or a species, what is rational from the viewpoint of the larger system need not be rational from the viewpoint of the agent. This may be important in trying to understand features of motivation and personality in human-like agents. For example a concern with reproduction and care about one's offspring pervades human personality (well most human personalities, if not all), and yet from the point of view of an individual the cost of reproduction and caring for the young is so high that it is highly irrational, especially for women.

4.3 A language for describing mappings

We need to describe mappings between regions of design space and regions of niche space. For example, designs need to be evaluated relative to niches, but, as already indicated:

(a) This will in general not be a simple numerical evaluation

(b) It will have to include descriptions of trade-offs, and possibly multiple coexisting partial orderings

(c) It may in some cases be related to evolutionary 'fitness' criteria, e.g. effectiveness in promoting survival and reproduction of a collection of genes, though other kinds of fitness will often be relevant.

(d) It will not in general determine unique design solutions for particular niches (as shown by biological diversity).

(e) It may have to include potential for future trajectories leading to a better fit between niche and design, either by individual development or learning, or by a succession of evolutionary stages.

All this may be a far cry from the current contents of AI books and journals. However, I expect it to be increasingly important over the next few decades, as more people come to understand the issues and grasp the shallowness and narrowness of much of current AI with its swings of fashion regarding particular mechanisms and architectures.

4.4 Resources and methods for exploring agent designs

It is commonplace for people in AI, and some of their critics, to make unnecessarily limiting assumptions about the methods, mechanisms or methodologies that can be used in AI.

1. AI can use whatever mechanisms will do the job: connectionist or neural mechanisms, chemical soups, or anything else. Restricting AI to use only a certain class of mechanisms would be like restricting physics to the experiments and mathematics available to Newton.

2. AI is not committed to the use of any particular class of representations. It is not committed to

the use of logic, or Lisp-like languages. It is part of the aim of AI to find out which formalisms are well suited to which purposes.

3. It has been fashionable in recent years to criticize 'toy' problems (e.g. simulated worlds). However, working on carefully chosen toy problems is part of a 'divide and conquer' research strategy, required for science. Controlled simplification helps in hard science. Using complicated robot eyes and arms does not necessarily cause one to tackle deep problems, as many frustrated students have found.

One important sort of simplification is to study what Bates et al. call 'broad and shallow' architectures, containing many functionally distinct components, each simplified. This may be one way of finding things out about some of the high level features both of design space and niche space (e.g. building systems and then discovering that they lack certain qualities that one had not previously realised were important). Even when we learn that certain architectures don't work, the study of why they don't work can be an important contribution to knowledge, and help us to a fuller appreciation of alternative designs.

Moreover, for us the broad and shallow approach is not a *permanent* commitment. It may help to clarify requirements for progressive deepening of the design. In some cases this can be done by incorporating what has been learnt via a narrow and deep approach. In other cases it may draw attention to what is missing from such an approach, e.g. a study of visual perception that assumes the sole function of vision is to produce information about the structure and motion of objects (criticised in Sloman 1989).

5 Expanded objectives

Our explorations are both scientific and concerned with practical engineering applications.

Scientific objectives include:

1. Trying to understand human capabilities

2. Trying to understand other animals

3. Trying to understand the space of possible designs and how they relate to different niches (capabilities, etc.)

4. Trying to understand which sorts of trajectories in design space and niche space are possible and under what conditions.

Too much AI work merely produces one implementation, without any analysis of the design, the region of niche space or alternative designs. However, this may suffice for certain engineering objectives.

Engineering objectives include:

1. Trying to design useful machines that can do ever increasing subsets of what can currently be done only by humans and other animals.

2. Trying to design machines that (for certain jobs) are better than humans or other animals. Often these are not 'stand-alone' machines but components of other machines.

3. Trying to make machines better suited to interact with humans (this depends on learning more about humans).

Other, less obvious, practical objectives include:

4. Developing new educational strategies and technologies: you can't improve human learning

without knowing how it works normally.

5. Understanding ways in which the human mind or brain can ‘go wrong’ may help us design better therapies. You can’t easily fix something if you don’t know how it works normally!

6. Designing new forms of recreation, new toys. (This depends on the scientific and engineering advances.)

A requirement for progress in all of this is the production of better tools, and that has been a constant feature of AI research. So we can add:

7. Meta-design: Designing new tools and techniques, including programming languages and forms of representation, to help with the process of exploring and implementing designs.

6 The Cognition and Affect project

The Cognition and Affect project at the University of Birmingham has been concerned with all the above issues for several years, although severely limited resources have forced us to concentrate our main efforts on tiny subsets of the task. The project has come up with some partial requirements for human-like designs, a preliminary partial specification of a type of architecture that might explain human like capabilities, and some preliminary attempts to map that architecture onto the phenomenology of common human emotional states, especially grief (Wright, Sloman & Beaudoin, to appear).

Our work has included the following:

• Collecting ‘requirements’ for human-like intelligence, such as:

1. The ability to handle multiple independent sources of motivation, some to do with physiological needs, some to do with the individual’s preferences, tastes and ambitions, and some to do with the needs of the culture,

2. The ability to cope with many concurrent processes (perception, plan execution, planning, thinking, etc.)

3. The ability to cope despite limited multi-processing capabilities for ‘high level’ processes.

• Exploring a variety of designs and design fragments, including:

1. Attention-filtering mechanisms to ‘protect’ resource-limited management processes.

2. Motive generation and reactivation mechanisms.

3. ‘Mood’ changing mechanisms.

4. Meta-management mechanisms that help to control the resource-limited management processes.

5. Aspects of a control hierarchy that accommodates both long term and short term change.

• Producing ideas that can influence therapies, e.g. for problems involving control of attention.

• Producing interactive demonstrations that might be used for teaching psychologists, therapists or counsellors. (So far only very primitive implementations exist.)

NOTES:

1. We are particularly interested in ‘broad’ architectures, so initially they are very ‘shallow’. Deepening can come later.

2. Often the process of working out an implementation reveals inadequacies in theories, long before there’s a running program to test! At present that’s the most important role of implementation.

3. It's inevitably a multi-disciplinary exercise requiring contributions from philosophy, psychology, psychiatry, neuroscience, biology, etc.

4. It's very difficult!

More information is available from our ftp site:

ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect

6.1 The Minder scenario

In order to focus our investigations and provide possibilities for useful implementation with restricted resources, we developed a specification for an extendable scenario in which to study some of the processes that interested us (Beaudoin & Sloman 1993, Beaudoin 1994).

The scenario involves a simple 2-D nursery (or creche) which has a collection of robot 'babies' (minibots) that roam around in various interconnected rooms, and a *minder*,[1] with a mobile camera and a mobile hand or 'claw'. The minder has to look after the babies, keeping them out of various kinds of trouble and rescuing them when they get into trouble, until they develop to the point where they can leave the nursery.

Types of problems the babies can encounter include the following:

(a) They can fall into ditches and die.

(b) Their batteries may run down, so that they need recharging, at a recharge point.

(c) If the charge gets too low, they die.

(d) Overcrowding can cause some babies to turn into 'thugs', which then have a tendency to damage other babies.

(e) Damaged babies need to be taken to the medical centre for repair.

(f) If the damage is too great the babies die.

The scenario can later be modified in many ways. The babies can either move around at random or act on goals with some degree of intelligence, exactly what sort of intelligence determines the niche for the minder. The scenario can either have a fixed set of babies, all to be kept alive till they are ready to be discharged, or a steady stream of incoming babies, so that the minder's task is to maximise the rate at which mature surviving babies can be discharged. The minder might be given an auditory sense as well as the camera, e.g. so that sounds of trouble can trigger visual investigation. Predators could make the task harder, and so on.

Initially there was only one minder with a movable claw and a movable camera with a restricted view (e.g. limited to one room at a time). The minder had no other embodiment, since our main task was to design its mind, and that was a big enough task. The camera could look at one room at a time and be moved around, either at random, or by cycling around systematically, or under the control of an attentive process driven by current goals. Alternative more complex scenarios are being investigated. We are also looking at even simpler agents in the context of evolutionary experiments, led by Riccardo Poli.

We chose a simple world, with as little complication in physics, perception, motor control, as possible, because that's not what we are interested in. We are interested in the mind, and the control of the mind, and this environment was designed to give the mind some hard control problems.

Later work could add a more complex body e.g. with auditory sensors, a more complex shape,

[1]In previous papers, we referred to the minder as a 'nursemaid'.

more parts for manipulating things, richer visual processing, and so on. Similarly the 2-D domain could be expanded to a 3-D domain, but that would enormously complicate problems that at present are not our main focus of interest. When the time comes we would expect to have to collaborate with a research group that has implemented a much richer 3-D world.

Another development is to have more than one minder, to introduce problems of communication and cooperation, both in planning and in acting. The babies could be given more intelligence and a wider range of personalities, perhaps even allowing some of them to learn to help with the minding (a possibility being investigated by my colleague Darryl Davis, who calls them 'minibots').

Even within a very simple 2-D world, the minding task can be made more or less difficult in various ways, e.g.

• changing the numbers of babies to be looked after,

• altering the relative speeds with which the babies can move and the minder's bodily parts can move,

• more interestingly, altering the relative speeds with which processes occur in the environment and 'mental' processes of various kinds occur in the minder, e.g. analysing perceptual input, generating new goals, evaluating the relative importance and urgency of goals, planning, etc.

The last is particularly important as one of our concerns is to see how requirements related to limited processing speeds for high level 'management processes' constrain the design of an architecture able to cope with asynchronous internal and external events. (See, for example, Simon 1967, Sloman & Croucher 1981, Sloman 1987, Beaudoin 1994, Wright et al. to appear).

The minder has a collection of different sorts of internal processes, all going on in parallel. Our specifications partly overlap the sorts of things described at the workshop by David Moffatt and Bryan Loyall. The internal processes include things like:

• realizing that there is a need to consider some new goal

• deciding whether to consider a particular goal

• evaluating a goal in various ways, e.g. importance, urgency, cost, likelihood of success

• deciding whether to adopt or to reject a goal

• deciding when to act on a goal: meta-planning

• deciding *how* to achieve a goal: e.g. by planning, or selection of an existing plan

• detecting and resolving conflicts between goals and plans

• carrying out a plan and monitoring progress

Some of these tasks turned out to have unexpected complexities. Urgency, for example, turned out to include both 'terminal urgency' of a goal which is a measure of how much time is left before it is too late, and generalised urgency which is a function from length of delay to costs and benefits. E.g. generalised urgency may be cyclic: the benefits may go down then up according to time of day, or the season (e.g. planting grain). (Beaudoin 1994).

It also turned out that there was no fixed sequence in which management tasks needed to be performed, so that a simple flow chart was not adequate. E.g. sometimes a goal can be evaluated and a decision made whether to adopt it prior to any planning, whereas in other cases at least partial planning may be required to evaluate costs and benefits and potential side-effects. (Beaudoin 1994 gives more details.)

These 'management' processes take time, and some of them take unpredictable amounts of time. Many of the problems would be simplified if the minder (or a person) could think with infinite speed. But our assumption is that in human-like agents there are reasons why processing resources, at least for a subset of the cognitive tasks, will be limited, and the amount of parallelism will be limited.

The limits to parallelism were first observed empirically and informally. Later we found a number of *design* factors explaining why it is to be expected that the amount of concurrency in management processes should be limited. The reasons are summarised in the appendix.

Our SIM_AGENT toolkit (Sloman & Poli 1996) was designed to support exploration of interacting agents with rich internal architectures in which we could vary relative processing speeds between objects and agents and between components within an agent. The toolkit makes it easy to change the speeds of sub-mechanisms within a simulation, both inside the mind and in the simulated environment.

One of our conjectures is that where some of the high level ‘management’ processes (such as occur in humans, though not necessarily in microbes or ants or rats) are limited in both speed and amount of parallelism, an extra level is required in the architecture, which provides what we call ‘meta-management’ processes, recursively defined as processes whose goals involve management or meta-management tasks. Because the definition is recursive we don’t need meta-meta-management mechanisms etc.

For example, a meta-management process might detect that a planning process is taking so long that an urgent goal will not be achieved, and decide to switch the management process to carrying out a partial plan in the hope that the plan can be completed later. Another might detect that the rate of occurrence of new problems is so high that switching between management tasks is preventing any significant progress. This could lead to raising of an ‘interrupt’ threshold for new goals or other information. Hence the dynamic attention filter in the architecture sketched below.

7 Towards a broad architecture for human-like agents

AI researchers cannot work on everything at once. Many rightly choose to work on narrow and deep mechanisms, e.g. concerned with vision, or planning, or learning, or language understanding. My main concern is how to put all those mechanisms together. So, like the OZ group at Carnegie Mellon University, I have chosen to concentrate on architectures that are initially shallow but broad, combining many sorts of functionality. Later research can gradually increase the depth, which would simultaneously involve increasing the complexity of the environment making the added depth necessary. We can also later increase the breadth, e.g. adding components to the architecture corresponding to evolutionarily older parts of the human brain that we share with many other animals, but which for now we are ignoring (e.g. the limbic system).

In any case, whether an architecture is a close model of any living organism or not, its study can contribute to the general exploration of niche space, design space and their relationships.

It is not possible here to give a full account of our work. So I’ll summarise some of the main assumptions regarding the sort of architecture we have been studying (though alternatives are also under consideration).

7.1 Automatic processes and management processes

It seems clear that there are different routes through human brains from sensors to effectors, from perception to action. Some of these routes seem to be shared with other animals, whereas others involve forms of cognition that may well be unique to humans, or at least restricted to a small subset of animals. The latter probably evolved much later.

Automatic, pre-attentive, processes

In particular, the older routes involve many automatic processes. These can be thought of as essentially being a large collection of condition-action systems which are probably implemented in neural nets, permitting a lot of parallel propagation of activation through the networks. The processes are 'automatic' in the sense that as soon as the conditions for some action occur the action (whether internal or external) is triggered.

Examples include low-level perceptual processing, posture control and many other processes triggered by perception, including internal perception of things like temperature changes, damage to tissues, the need for food or liquid, and other body states. These pre-attentive processes can trigger other pre-attentive processes.

Some of them can generate output, both in the environment (e.g. reflex actions, trained responses) and also internally, e.g. controlling internal states, such as generating new desires and driving learning mechanisms.

Attentive management processes.

Other routes from perception to action, at least in humans, include processes that are not automatic in the following sense. When triggering conditions for an internal or external action occur the action does not automatically happen. Instead alternative possibilities are considered explicitly (e.g. doing A or not doing A, using this plan or using that plan) and then a choice made between them. Sometimes very elaborate temporary structures (e.g. new possible plans) have to be created and evaluated as part of this process. In addition arbitrarily complex sets of previously stored information may need to be accessed and derivations made, combining old and new information. In general these processes involve combinatorial search: attempts to find combinations of ideas, or actions that will enable some problem to be solved or task to be achieved. Thus, there is not only selection between complete alternatives: many fragments of a solution may require choices to be explicitly constructed and selections made.

The mechanisms forming the implementation for the attentive management processes may themselves be automatic pre-attentive mechanisms. Something has to work automatically or nothing could ever get started.

Besides the functional differences just described it is possible that management processes and automatic processes use different kinds of representations and different sorts of control mechanisms. For example some neural nets provide mechanisms for mapping inputs in one space to outputs in another space, where both spaces have fixed dimensionality and well defined metrics. This could be very useful in automatic processing, though not so useful in problems requiring creation of novel structures of varying complexity.

The distinction between management processes and automatic processes is indicated crudely in Figure 2 (due partly to Luc Beaudoin and Ian Wright). The distinction is neither sharp nor very well defined yet. We cannot have good concepts and distinctions until we have a really good theory. In deep science, concepts and definitions come after theory. In shallow science, we often start with operational definitions so that we can get on and measure things, instead of thinking, which is

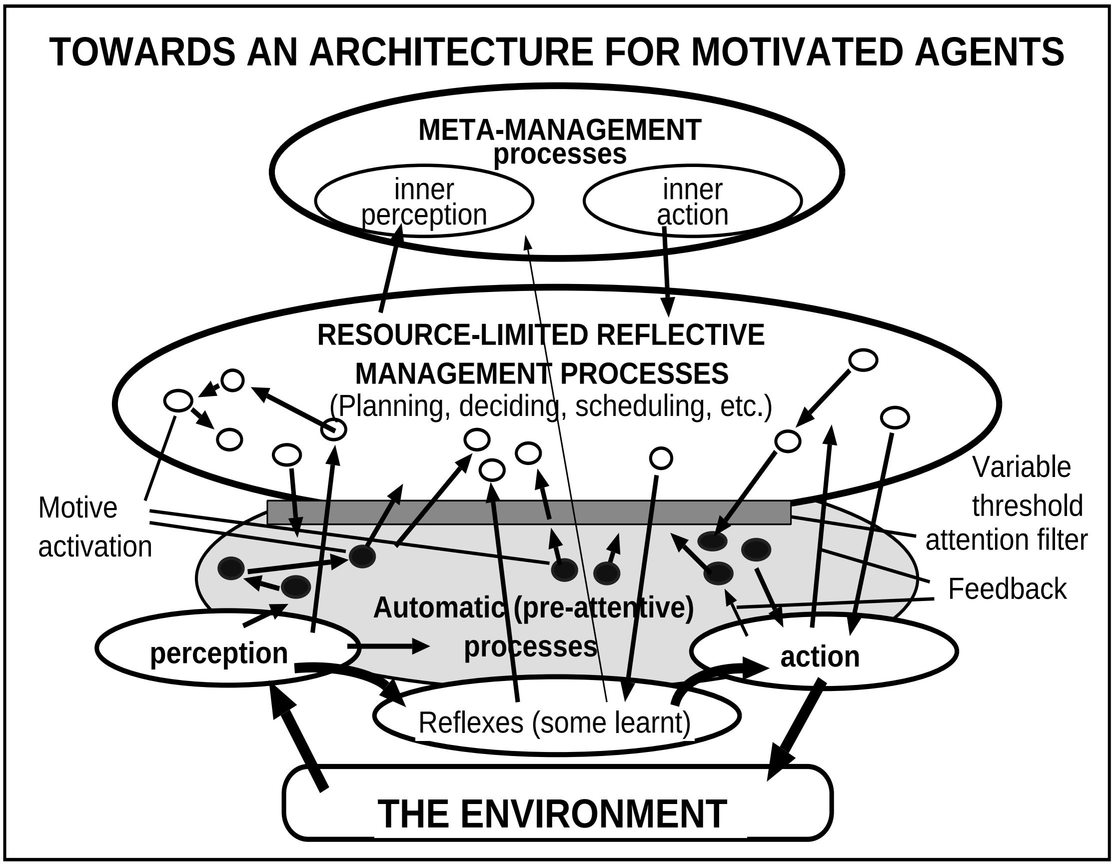

Figure 2: **Towards an Intelligent Agent Architecture**
There are several layers of control, involving different routes through the system, from perception to action. Some evolved very early and are shared with many other organisms. Some are newer, and less common.

much harder. So in what follows I am using provisional terminology that is part of a boot-strapping process of theory building and concept formation.

It seems to be a feature of management processes in humans that they are resource limited, unlike the automatic processes. It is as if the automatic processes have pre-allocated, dedicated, portions of the brain and they can operate whenever they need to, whereas different management processes have to share and re-use relatively slow mechanisms, so that parallelism is restricted, as explained in the appendix.

For example there are different kinds of mental tasks we can perform e.g. counting to oneself, reciting a poem to oneself, counting backwards from 10,000, singing a song silently to oneself, and so on. But we are not able to do several of them fluently in parallel, even though we can alternate between them, and even though there are some things we can do in parallel, e.g. hopping, clapping our hands and talking.

Because both the amount of parallelism and speed of management processes are limited, an interrupt filtering mechanism may be required, to prevent disturbance when the management current task is both very important and very urgent, or likely to fail if temporarily disrupted.

Although some new goals are automatically generated and automatically acted on in the ‘older’ part of the brain, concurrently with management processes, some tasks (e.g. those requiring planning or investigation) need to be handled by the management system. If they are both urgent or important, this may require interrupting and suspending or aborting some other activity.

Since such interrupts are not always desirable, one way of limiting their potential damage is to have an interrupt filtering mechanism with a dynamically varying threshold. This requires the automatic mechanisms to assign some measure to new goals which will determine their ability to get through the filter. I have called this the *insistence* of a goal. Insistence is one among many features of goals and other information structures controlling processing.

7.2 Processes involving motivators

There are many different sorts of processes involving motivators. Motivators are information structures with the potential to initiate or modify internal or external actions, either directly (in the case of innate or learnt cognitive reflexes) or as a result of processes of explicit evaluation of the motivator, acceptance of it, formation of a plan, and plan execution. The following seem to be among the internal behaviours concerned with motivators, which synthetic human-like agents with personalities will require.

- Motivator generation and (re-)activation and setting ‘insistence’ (interrupt capability).
- Mechanisms to suppress or ‘filter’ motivators, to protect resource-limited management processes.
- Management of motivators can include the following.
    - Assessing motivators. e.g. importance, likelihood of satisfaction, cost of satisfaction, urgency.
    - Deciding: whether to adopt the motivator, i.e. form an intention.
    - Scheduling: when or under which conditions to execute a motivator.
    - Expansion: deciding how to execute a motivator (planning).
    - Predicting effects. This can occur as part of evaluation or planning or other processes.
    - Assigning an ‘intensity’ measure. This is not the same as insistence: it influences the ability of the motivator to maintain control once it has gained control.
    - Detecting conflicts between motivators.
    - Detecting mutual support between motivators.
    - Setting thresholds for the management interrupt filter.
    - Termination of motivators. E.g explicit termination on satisfaction, or decay.
    - Detecting the relevance of new events to existing motivators.
- Meta-management: I.e. processes that (recursively) control management or meta-management processes (e.g. deciding which to do when).
- Execution of plans, with or without high level management.
- Learning: improving or extending performance, improving methods for assigning insistence, for assessing urgency or importance, for choosing in cases of conflict, etc.
- Extending the architecture: developing new abilities, or new ‘cognitive reflexes’.
- Global switches or modulators: e.g. mood changes, arousal changes, e.g. becoming optimistic and bold, or turning pessimistic and cautious.

7.3 Representing motivator structure

In order to be able to engage with all these different kinds of processes, motivators need a rich structure. They often include the following components, though they may have other specific features also. Some of these will vary over time, whereas others define the motivator and are fixed.

(1) Semantic content: for example a proposition, P, denoting a possible state of affairs, which may be true or false

(2) A motivational attitude to P, e.g. 'make true', 'keep true', 'make false', etc.

(3) A rationale, if the motivator arose from explicit reasoning.

(4) An indication of the current belief about P's status, e.g. true, false, nearly true, probable, unlikely etc.

(5) An 'importance value' (e.g. 'neutral', 'low', 'medium', 'high', 'unknown'), importance may be intrinsic, or based on assessment of consequences of (doing and not doing).

(6) An 'urgency descriptor' (possibly a time/cost function).

(7) A heuristically computed 'insistence value', determining interrupt capabilities. Should correspond loosely to estimated importance and urgency. This is used only for attracting attention.

(8) Intensity – determines whether a motivator that has already been attended to (thought about, acted on) will continue to be favoured over others that may be considered. This gives motivators a kind of momentum.

(9) Possibly a plan or set of plans for achieving the motivator.

(10) Commitment status (e.g. 'adopted', 'rejected', 'undecided')

(11) Dynamic state (e.g. 'being considered', 'consideration deferred till...', 'nearing completion', etc.)

(12) Management information, e.g. the state of current relevant management and meta-management processes.

In most animals, as in current robots and software agents, motivators probably have a much simpler structure. We need to explore the possibilities for a variety of different types of motivator structure. These will require differences in motive generation, in management processes, in meta-management processes and in execution processes.

There may be individual differences among humans too.

Exploring 'design space' will help to show what is possible.

8 Deepening the design: visual perception

Figure 2 is in some ways misleading as it suggests that components of the architecture have a simple structure. In particular, boxes concerned with perception need to be far more complex than the figure indicates. Attempts over many years to model visual perception have suggested that at least for a human-like visual system something with the sort of complexity indicated in Figure 3 may be required.

For example, perception is not just a matter of recognizing patterns in the sensory input. Different levels of analysis are required. In the case of speech this is very obvious: besides the acoustic signal there are levels of interpretation concerned with phonetics, word recognition, syntax, semantics and what the speaker intends to achieve (sometimes called 'pragmatics').

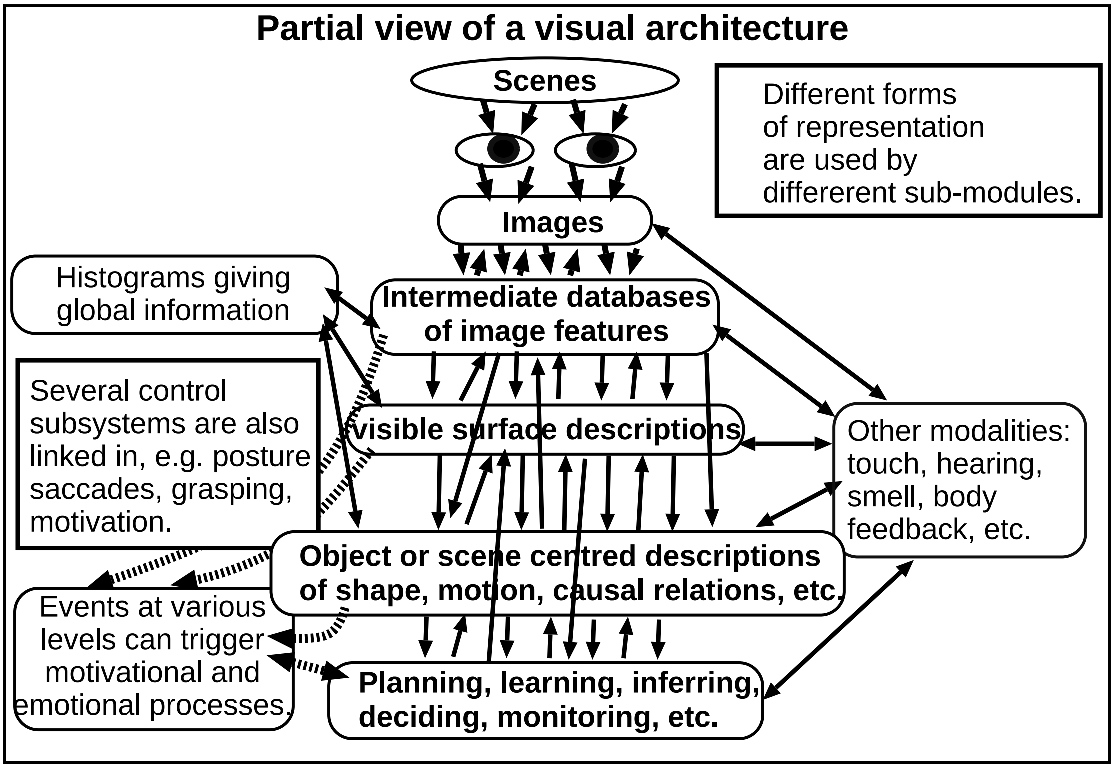

Figure 3: Sketch of a visual sub-architecture

In the case of vision there is often, though not always, a requirement to go far beyond the 2-D structure of the retinal image or the optic array and find the 3-D structure of the environment (Marr 1982) and many non-spatial properties and relations of objects including their causal relations and potential for change, what J J Gibson referred to as 'affordances' (Sloman 1989). For example, I need to be able to see not only that a surface is flat and horizontal, but also that I can sit on it and it will support me and I need to see a cliff as potentially dangerous.

Besides the internal complexity and variety of tasks in visual processing mechanisms, they also have many links to other sources of information besides the visual images. For instance visual processing can be influenced both by input via other current sensors (what is heard or touched) and also by general knowledge about the type of thing being seen. Moreover, instead of having only one form of output, descriptions of spatial structure and motion, as Marr suggests, visual mechanisms may also have many outputs of different sorts to other sub-mechanisms, including posture control and some motive generators. Some of the cross links merely transmit control signals, whereas others transmit information about contents of intermediate databases, or about the environment.

Vision is a primary example of the claim that there are myriad routes through the system, including many routes through the automatic processing mechanisms serving many different purposes, including controlling breathing, heart rate, posture, sexual arousal, various sorts of attraction or disgust, and no doubt many things we don't yet know about.

Visual input is only one modulator of posture. Some aspects will be partly a result of physical structure, e.g. congenital deformities, or a person's height relative to most doorways or other individuals. Other aspects could be due to early childhood experiences, e.g. having a brutal, easily provoked, parent might lead to the development of a very retiring and diffident posture.

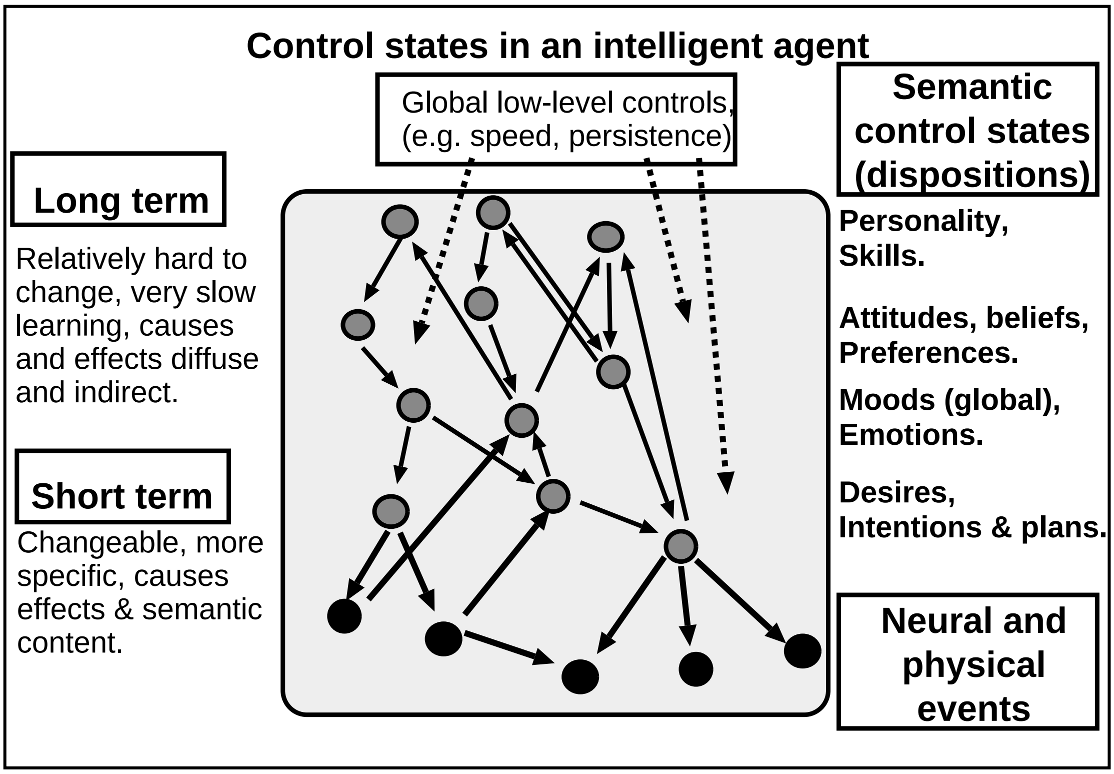

Figure 4: Control states of various kinds: see text.

Visual processing that triggered reminders of unpleasant interactions could influence posture, e.g. producing a tendency to cower, in parallel with more common posture control signals.

Much of the detail concerning what sort of processing occurs and which cross links occur will be very dependent on individual experience, and will form part of a unique personality. An architect will see a building differently from its occupants. Perception of spiders affects different personalities in different ways.

So personality is not just a feature of some high level control mechanism, but is distributed throughout the whole system. Moreover, there isn't *one* thing that is distributed - personality is multi-faceted as well as distributed. That is one of the reasons why it is so hard to change: there is so much to change.

9 Control states in an intelligent agent

I have tried to indicate this in Figure 4, which is an impressionistic diagram intended to suggest dynamic aspects of the human-like architecture, namely existence of different levels of control, with varying types of influence, different life-spans and different degrees of ease of change.

The dark circles represent an event stream, whereas the unfilled circles represent control states, some long term, some short term, some specific and goal-directed (such as desires), some more global (such as moods). Some of the events are physical, e.g. being struck by a falling apple, others mental, e.g. being struck by a thought. Control states are subject to some influences that are bottom-up (event driven) and others that are more top down, e.g. the influence of a longer lasting or more general control state, including aspects of personality.

A very sudden and intense pain in the hand may produce very specific and short term manifestations, e.g. movement away from the source of the pain. However, a very long lasting intense pain might cause all sorts of complex processes, including going to a surgical specialist, taking drugs, reading about neuroses, etc. (Are such differences in effects produced by similar mental states also found in other animals?)

Some of the mental events and enduring states have semantic content (e.g. attitudes, desires) while some (e.g. moods, states of arousal) merely modulate processing locally or globally, e.g. speed, risk-taking, alertness, amount of deliberation prior to action. The modulation may be quantitative (e.g. speeding up, slowing down) or structural (e.g. switching from one strategy or language to another).

Arrows represent causes of differing strengths and differing time-scales, some deterministic some probabilistic. Various routes through the system are causal influences linking events and enduring but modifiable states.

The distinction between automatic and attentive processes can be made at all levels in this diagram.

Unfilled circles at lower levels in the diagram are supposed to represent control states that influence events fairly directly, and which are typically also readily changeable from one moment to another, e.g. how thirsty one is, or whether somebody has annoyed one so much that one just wants to walk away.

Higher up the system are things which are more indirect in their influences and also capable of having a broader range of influences, for instance tastes in food or music, or political attitudes. These can influence events only indirectly, via changes to dispositions lower down the system - e.g. creation of new motivators.

A control state is *dispositional* insofar as it can exist without producing any actual change in the event stream, until the conditions are right for it to do so. (This was one of the main themes of Ryle 1949). For example, the brittleness of glass is capable of determining how it breaks, but can persist for a long time without producing any particular manifestation, until the glass is struck hard. Similarly many dormant mental states produce no effects until the conditions are right.

Some of the 'higher level' states also have a very general range of potential influences, whereas others are more specific. For example, a state of nervousness that controls my current posture is very specific. By contrast, my preference for old-fashioned movies as opposed to new ones will not directly influence specific items of behaviour. It might sometimes get me to board a train to go and see some movie I have heard about. At other times, it might make me look closely at something, e.g. a poster advertising a film show. At some other time, it may make me go to a library to read a book.

High level control states can evolve. For instance tastes in films may develop over time. These changes need not be manifest. A high level dispositional mental state S may be changed in such a way that if some new information comes up later S will produce behaviour different from what it would have produced if it hadn't been influenced earlier. yet the change need not actually be manifested for a long time, if ever.

The fact that some dispositions are never manifested, because their activating conditions do not occur, or because other things suppress their influences does not mean that they do not exist. They are merely dormant.

Moreover, internal influences, when they do occur, need not lead to external signs. People with a lot of self-control often have states that do not manifest themselves externally, whereas others reveal nearly everything. This is another aspect of personality: how much and what sort of control exists over expression of current thoughts, desires, feelings, etc. This can have an important effect on social relations, via complex social feedback loops.

These differences between different levels in the control hierarchy do not correspond in any direct way to the difference between management processes and automatic processes depicted in Figure 2. Differences in level of the sort discussed in this section can occur in both the management processes and the automatic processes. Understanding these issues requires far more exploration of design space and types of control architectures.

10 Problems in defining 'emotion'

One manifestation of personality is the variety of emotional states an individual is capable of. However, discussion of that is bedevilled by the fact that the word 'emotion' has such a wide variety of uses. There seem to be as many different definitions of 'emotion' as there are people claiming to be studying emotions.

At one extreme are people who write as if every motive is an emotion, if it's a strong motive. For instance, for such a person a strong desire for promotion would be an emotion. I'd call it a motive.

Another kind of extreme view defines 'emotion' in terms of particular mechanisms of the brain, e.g. the limbic system. This focuses attention on a concept of emotion that can apply both to humans and many other animals with whom we share an ancient evolutionary history.

Others define 'emotion' in terms of external behavioural manifestations, e.g. whether the corners of the mouth go up or down, whether tears come out of the eyes, whether an aggressive or cowering posture is produced, and so on.

Some define it in terms of the things which are at the boundary between internal and external processes, e.g. muscular tension, proprioceptive sensory information, galvanic skin response, and so on.

In my own work, I have concentrated mainly on a subclass of emotional states that seem to be common in humans but may not occur in all other animals. These are states that deeply involve high level cognitive processes and which are often socially important, for example grief, anger, excited anticipation, dismay, pride in achievement: the kinds of emotions that poets and novelists write about.

These are all states in which there is a partial loss of control of high level cognitive processes. These are clearly included in common non-technical uses of the word 'emotion', but to avoid confusion with other technical definitions our project often uses the word 'perturbance' to label such phenomena. The word is intended to resonate with the concepts of 'perturbation' and 'disturbance'.

10.1 An example of perturbant emotional states

I once heard a woman interviewed on the radio who was talking about her grief. Her teen-age son who had been killed in a road accident a year before.

She did not talk about the state of her limbic system, nor about whether her muscles were tense or not, nor her facial expressions, or her posture, nor whether she was sweating or not. What she talked about was her state of mind.

One of the themes that kept coming up was the extent to which she was unable to control her own mental processes. She was constantly drawn back to thinking about this child, about what he would have been doing now, about whether she could have prevented the accident, about whether to keep his room as it was when he died, as opposed to letting one of the other children move in and use it. Her inability to control what she was thinking about meant that she could not pay attention to normal tasks and carry them out properly.

She found this inability to get on with her life very debilitating. It made her miserable, and it affected other people also. She could not do most of the other things she was supposed to do, including looking after her children. And that made her very depressed.

She was grieving, and she was depressed, and she felt guilty about her inability to cope and she desperately wished her own life could end somehow, though she did not wish to take it herself. So she had several emotions piling up on top of one another.

The possibility of several different sorts of mental states occurring simultaneously, is one of the kinds of consequences of the type of architecture I have been talking about, whereas some theories of emotions claim that only one can occur at a time. The links between the architecture and states like grieving is discussed more fully in (Wright et al. 1966).

Grieving is not the only case. I suspect most people have experienced such states. For instance, a person who has been humiliated by someone else, or who made a fool of himself in public may remain concerned about it for some time after, not in the sense that they simply wish it had not happened (which seems to be Frijda's sense of 'concern' in Frijda 1986), but in a stronger sense: it holds their attention. They can't put it out of their mind. They are drawn back to it, even when they don't wish to be.

Anger is another common example. Consider someone who has done something that you did not like. It stopped you in achieving your goal. You may wish he hadn't done it. You may wish to do something in return to stop him doing it to you again. However, merely desiring revenge is not yet the sort of perturbant state I am talking about, though it is an aspect of anger. In addition you may find that you are partly out of control. The desire to get your own back is something you can't put out of your mind. It continually tends to draw your attention away from other things. However that is a dispositional state, and like other dispositions it may temporarily be swamped by more powerful influences, e.g. seeing a child in sudden danger.

I suspect that such states cannot occur in most other animals. Although they may share the states and processes that occur within what I have called the 'automatic' part of the architecture, they do not have the sort of architecture that makes it possible for them sometimes to be in control of their thought processes and sometimes not. A rat may sometimes be terrified, but can it lose control of thought processes if never has control?

10.2 Why do perturbant states occur?

The possibility of perturbant states is inherent in the sort of architecture I have been discussing, though not because the architecture evolved in order to produce perturbant states. Similarly the possibility of thrashing is inherent in many multi-processing computer operating systems, but not because that is something intended by the designers. The possibility *emerges* from other functional aspects of the design.

In this case we have management mechanisms whose parallelism is limited, and consequently they need to be protected by some kind of filtering mechanism. I've suggested that new motivators (and other information items) may be assigned a level of insistence which determines their ability to get through the filter and interrupt management processes. Because the insistence level is assigned by *automatic* processes it may not be possible to prevent a high level of insistence being assigned to goals that have already been rejected, or which conflict with some high level objectives.

When attention is frequently interrupted and held by items that have been assigned a high insistence level which is *inconsistent* with goals or strategies selected by the meta-management processes, then the agent has partly *lost control* of his or her thought processes.

A full account of such perturbant states would show how they are dispositional states that can persist even though temporarily dormant because other more insistent and more acceptable thoughts and goals grab and hold attention (Sloman 1987, Wright et al. 1996).

11 Conclusion

This exploration of niche space and design space and their relationships is a multi-disciplinary affair, and includes not only: exploration of various architectures, mechanisms, formalisms, inference systems, and the like (aspects of natural and artificial designs), but also the attempt to characterize various kinds of behavioural capabilities and the environments in which they are required, or possible.

I do not believe we can understand human personality except in the context of a study of possible designs for human-like agents in human-like niches. This study is also a prerequisite for constructing synthetic agents with aspects of human personality, attitudes, emotions and motivation in synthetic agents. In this and other papers I have tried to show how such features may be explained on the basis of an underlying information processing architecture and its design requirements.

The implications of such a study are profound: e.g.

- for engineering,
- for biology,
- for psychology,
- for philosophy,

and

- for our view of how we fit into the scheme of things.

Acknowledgements

This work was supported by the UK Joint Council Initiative, the Renaissance Trust, and a grant from the University of Birmingham.

I am grateful to Robert Trappl and colleagues at the Austrian Research Institute for Artificial Intelligence. In particular Gerda Helscher prepared the transcript of my talk and the discussion from the audiotapes.

Thanks to Luc Beaudoin, Ian Wright, Tim Read, Christian Paterson, Chris Complin, Riccardo Poli, Darryl Davis, all past or present members of the Cognition and Affect project at Birmingham. Additional papers by the group can be found at the ftp site:

ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect/0-INDEX.html

or via this Web page

http://www.cs.bham.ac.uk/~axs/cogaff.html

APPENDIX: Discussion

What follows is an edited transcript of the discussion leading on from the talk. Some parts of the tape were very unclear and some of my editing may be wishful thinking. The transcript has not been checked with other participants, and as I do not have access to the original audio tapes I may have taken some wrong decisions regarding ambiguous or incomplete portions of the transcript. I have moved some portions of the discussion back into the main paper, where they seemed most relevant.

A.1 Metric for success of the design

BLW: What's the metric of success for the design of the minder?

Sloman: At the moment, we don't have any well-defined notion of success for our experiments, but you could easily use a measure of how many babies the minder manages to keep alive over a period of time.

For the existing simple implementation, we can vary the relative speeds of (a) movements of the babies (b) movement of the minder's hand, and (c) internal processing in the minder. We then find that, for instance, for some settings, it's impossible to keep even more than two babies alive, and for other settings more babies can be kept out of trouble, though that depends on the initial geographical distribution of the babies and their behaviour, which has random elements.

Another sort of evaluation, which would be appropriate when we have a full implementation including the meta-management processes, would be to see what happens when we make external events generate new goals faster than they can be handled by the management processes, and then see whether introduction of meta-management processes improves success at keeping the babies alive.

A.2 Physical and other resource limits

Several points of clarification emerged regarding the claim that management processes were resource limited.

BLW: ... contrast with automatic processes that you could do, many of them at one time?

Sloman: That's the point. There's, as far as I know, no particular reason why parallelism should be limited in the pre-attentive parts of the architecture. Whereas physical limits, such as having only two hands .....

BLW: That's right. That's right. That's all I was getting at. There are physical resource limitations.

Sloman: Yes. On the other hand, for a distributed robot, even that might not be the case. You might have a robot with many hands, like an octopus, doing many things at once, which would have fewer physical limits to parallelism than we have.

But I am not talking only about physical resource limits. I am talking about limitations of information processing resources.

An interesting question is whether it is just a fact of how we evolved that we have these resource limits, or whether there are good design reasons for this. (That's always a good question to ask whenever you discover a fact about human psychology.)

In this case, I think there are good design reasons for limits to parallelism.

One of them has to do with the fact that if you want to learn from what you do, you had better not do too many things at once, or you will have to solve a complex credit assignment problem, i.e. deciding which combination of what you did was responsible for the good or bad things that happened unexpectedly. The number of possible combinations is an exponential function of the number actions, and can therefore rapidly become intractable, especially where effects can be delayed.

Another type of limit may arise from the need for a long-term memory store, which is content-addressable and makes use of a highly parallel associative engine, but can only be given *one* question at a time to answer.

Another type of limit may be concerned with the requirements for building temporary structures of unbounded complexity, during planning and problem solving.

Another may be due to the need to have a single coordinated process at the 'top level' to reduce the risk of independent sub-processes taking mutually inconsistent decisions (Sloman 1978, chapters 6 and 10).

There may be other reasons, design reasons not accidents of history, why sophisticated agents are limited in the number of 'management' tasks that they can do simultaneously.

The more automatic processes are not limited in their parallelism because they don't access common information stores, and they don't do sophisticated learning, and they don't have to build temporary structures of unknown complexity in re-usable workspaces. Notice that I am not saying that automatic processes never build complex structures: visual and language understanding processes almost certainly do. It may be that one effect of certain forms of training is the allocation of dedicated portions of the automatic mechanisms to tasks which thereafter can run in parallel with higher level processes. Learning to sight-read music is a spectacular example.

When the circumstances are appropriate, the automatic, mechanisms just get on with their tasks, with minimal mutual interference.

The meta-management processes, that I mentioned earlier, have to do with working out what to do next within this resource limited system. And one difference between different architectures might be whether meta-management is present or not.

If we had good ways to study mental architectures, we might find out which organisms do and which do not have architectures of the types discussed here. I suspect we will find many organisms that share with us only what I have called the automatic processes. But my impression is that very few species also include the attentive management processes and the meta-management processes for controlling them.

Perhaps even birds that perform complex nest-building tasks have only automatic mechanisms that respond to triggers and produce stereotyped responses, unlike a human house builder who sometimes explicitly considers alternative structures, alternative materials, and alternative construction sequences before acting.

Perhaps only a few other species, e.g. bonobos, chimps and nothing else, have such meta management processes. Perhaps only humans have them.

So, this whole approach defines a research programme for biologists.

A.3 How is filtering implemented

BLW: Can you say a little bit more about internal perception and internal action, or –

Sloman: Preferably not right now. Ok? Because there is just too much.

By the way, David Moffatt mentioned the filtering mechanism, which his system doesn't need, though I have claimed it is needed because some new goals need to be able to interrupt management processing whereas others should not.

Now, I have a totally open mind as to how that's implemented. It might be implemented by neural nets, which don't have an explicit filter, but just allow certain subnodes to become active under some conditions and not under others, because of the operation of excitative and inhibitory links.

Alternatively there might be a separate mechanism with an explicit threshold set by higher level processes.

The important point is not how it's implemented, but the sort of function that is needed within the total architecture.

BLW: Can you try .... to explain the requirements for the nursemaid project, characterizing that as some kind niche that will then require certain kinds of design to give us a sense of what you are doing.

Sloman: At the moment, the niche for the minder is rather simple. But it has a number of features which were chosen to drive the research.

One is that the world has a number of different things happening independently, which are capable at any time of generating new goals. So it's not like a system where you have a 'user' giving a command, and the system makes or selects a plan and after completing it comes back for the next command, like a simple computer interface. The minder, like a typical windowing interface, must have asynchronous perceptual monitoring of the environment, concurrently with everything else that's going on inside it, e.g. planning, problem solving, etc. So, that's an example of a requirement that forms part of the niche.

Another requirement is that the speed at which things happen in the environment should be such as to make these internal resource limits significant, so as to show up the need for an internal architecture that gets round the problem of the resource limits.

What that means is that you can't have a system which deals with each new goal by immediately investigating fully what its potential consequences are and what the costs and benefits of achieving it are, and deciding whether to adopt it, and if so when and how it could be achieved, and so on.

That is not possible because such processes could interfere with other important current planning or deliberation processes. This would not be a problem if management mechanisms supported unlimited parallelism or had no speed limitations, or if things happened so slowly in the environment that all internal processes could run to completion in any order without opportunities being missed.

Another requirement was that the environment should be unpredictable. That means that the system cannot decide in advance when to interrogate the environment and build that into its plans. That is why asynchronous concurrent perceptual processes are needed.

Gap in transcript due to tape change.

In a slightly more sophisticated simulated world, the minder might be aware that something over there is crying for help, but it does not know what or why, or where exactly. So it now has a further task, which is deciding whether to go and find out exactly what the problem is. This contrasts with the simplest version where the minder always knows exactly what the problem is, and its management task is deciding whether to act on this problem or another problem, or how to achieve particular goals, e.g. helping a baby with low charge recover by taking it to the recharge point, or picking up and moving a baby that has got too close to a ditch.

The current domain gives the minder a small collection of requirements to be satisfied, subject to trying to keep as many babies out of trouble as possible. Those requirements already make our design problem non-trivial, but we can go on adding more requirements. For instance we could add the requirement to be able to cooperate with another minder, in order to be able to do things sensibly. For instance, if there are two minders and two babies are in trouble, the minders shouldn't each go to the baby nearest to the other minder.

A.4 Comparison with real-time scheduling systems

??: ...(??)...that's the kind of problem that's solved better by a computer dedicated to doing real-time scheduling, and not by person.

So, what is it about this problem of trying to schedule these different tasks, that is so particular to humans?

Sloman: If you have a computer of infinite speed, there will be infinitely many different designs that will all be functionally equivalent. Ok?

??: Ok, you have some optimal conditions of performing these various tasks?

Sloman: I am not defining a notion of optimality. I am assuming that there is going to be a certain minimal architectural structure, because I think that's how we (humans) do things. I then try to see how to make that work.

But I am perfectly open to someone else deciding to explore different sorts of assumption from the sort we are talking about, or even running genetic algorithms to see what would evolve naturally in a whole variety of situations. Maybe we will find a collection of different solutions, or solutions which work in different ways.

One of the things about the mapping between niche space and design space, which was on the slide is that there are different styles of arrows, because there isn't a simple notion of a design fitting or not fitting a niche. There may be different ways of fitting and many trade-offs.

You may have something that works very well under normal circumstances, but perhaps a slight change of circumstances makes it go wrong, and other designs which don't work as well in normal conditions, but work better in abnormal conditions. Similarly one design may be better as regards speed or precision, and another may be less costly in fuel, or easier to build.

So, there is a whole variety of different kinds of issues. And I suspect we don't yet understand what all those issues are. So, we have to explore that. That's a partial answer to your question about alternative solutions.

A.5 Predictability and 'cognitive friendliness'

BP: Did I understand you right that you said the environment should be unpredictable?

Sloman: Not that it *should be*. It *is*. And we want to try to understand what the implications of being able to deal with an unpredictable environment are.

BP: Are you saying this environment (i.e. the minder scenario) is unpredictable?

Sloman: Yes. The minder does not know what each baby is going to do.

BP: Yes. But you have a lot of these constraints. You know that they won't be in that other room.

Sloman: It's partly predictable. Yes. Every environment is partly predictable. In fact, I believe that is in part a result of co-evolution between us and our environment.

BP: I think that this is what makes you able to live at all, these persistent features of the environment.(??)

Sloman: Yes. I have a concept of the cognitive friendliness of the environment. There are different dimensions of cognitive friendliness.

Some of them have to do with the availability of detailed information. So, for example, if we didn't have all this electromagnetic radiation bouncing off the surfaces of things onto detectors in our retinas which are very well tuned to those frequencies, the environment would be cognitively less friendly to us than it is.

Another type of cognitive friendliness concerns the amount of information and the complexity of processing required to discriminate objects, and whether the physical structures of objects that are importantly different in their behaviour are hard to differentiate. For instance, suppose some trees produced edible fruit and others produced poisonous fruit and the only way to tell which was which was to count the leaves on the whole tree to find whether the number was odd or even. That would be cognitively unfriendly.

But by and large, things evolve so that this sort of case does not occur. So, yes, you are right, our environment not totally unpredictable.

A.6 How many levels of management

BLW: (transcript unclear) raised a question about the meta-management processes being defined recursively, and how many levels of management might be required.

Sloman: In practice, the recursion probably runs out after two or three steps.

BLW: Yes. So one got this feeling that you have, that that is sufficient independent of the niche characteristics. Is that because you have performed an exploration of a large number of what niche spaces might be? Or just sort of a priori –

Sloman: It seems to me to be an empirical fact that we have something like this meta-management ability. Up to a point it works. However, I would not say it's *sufficient* in any general sense, because it doesn't solve all the problems.

BLW: So, you think that the niches that people actually encounter seem to motivate these three levels and nothing more. But an exploration of more different niches might lead to additional aspects of architecture.

Sloman: It might. But some of those additional aspects could require more concurrency at management levels, which I have claimed we don't have, for reasons explained previously.

Thus limits to concurrency could rule out coexisting processes involving management, meta-management, meta-meta-management, etc.

But of course, if you build a robot system, you might be free to put together whatever mechanisms you like, and you might find a totally different architecture, which, for instance, was much better than the sort of mechanism I have discussed. The control system required for a complex automated factory might be an example of a different sort of architecture.

So, I regard it as an open question, whether certain niches which are in some ways like ours, but in some ways different, would require very different architectures. And maybe even for niches like ours, there may be different successful solutions – different sorts of mind.

A.7 Why filtering is needed

??: So, the meta-management was supposed to identify processes that are identified with control, self-control.

Sloman: Processes that control management or meta-management.

??: So, this is your explanation for human emotion being –

Sloman: Oh, I haven't said anything about how meta-management relates to emotions.

I previously talked about the grieving mother, not being able totally to control her thought processes. And, that's linked to the fact that we need the attention filter, or something equivalent to a filter, which means that not every new *potential* user of high-level resources will automatically get those resources.

We can imagine designs which once they have started to attend to a problem are never diverted from it until that problem has been solved. But such single-mindedness could be disastrous in cases where something more important and more urgent turn up. So new motivators should be able to capture attention and re-direct management processes sometimes.

However, we can't just let every new motive have that effect, because the management processes have limited parallelism and some tasks are urgent and important and intricate, requiring full attention to all details. I don't want the brain surgeon who is operating on my brain to be distracted by every sound or thought about whether to have a holiday in Spain. Similarly if someone is giving you intricate instructions on how to defuse a deadly bomb, you had better listen all the time, because you might miss some crucial information otherwise.

So, we want to allow the possibility of interruption of high-level management processes, but we also sometimes want to prevent that. And that's where this notion of dynamically variable filtering mechanism comes from.

Exactly how it varies is an open question. I originally started with the idea that that a simple one-dimensional threshold might be enough, but Luc Beaudoin, in his PhD thesis suggested that instead of a simple numerical filter something more qualitative would be required, so that, for example, the whimpering of a baby, however quiet, could divert attention even when much louder noises from the kitchen could not. The type and level of threshold would then be determined by meta-management processes that determine whether the current management tasks should be easily interruptable.

A.8 Insistence levels are unreliable

The automatic processes that generate new potential distractors and assign insistence values to them cannot themselves be fully intelligent, because they would themselves then need all the resources that the filter mechanism must protect from disruption. So the insistence assignment mechanisms must use 'quick and dirty' heuristics which give an approximate estimate of importance and urgency, perhaps based on previously learnt patterns. So they will sometimes get things wrong, allowing dysfunctional types of interrupts to occur.

How insistence assignments occur and what sorts of filter thresholds are set up in various kinds of circumstances will be yet another way in which personalities can vary.

It seems that many – not all – but many of the things that we call human emotional states are cases where something continues to grab attention, whether you want it to or not. Now, sometimes you don't mind because there is a kind of pleasure in being partly out of control. E.g. sometimes being passionate thought to be a good thing. Or people put themselves on roller-coasters, where they get into a situation where they are partly out of control, and they do it freely. There are all kinds of strange things people like to do to themselves.

So, I am claiming that the aspect of emotional states which I call perturbance, which has to do with management processes being partly out of control, can occur only in the context of an architecture with sufficient richness to support the possibility of management processes being controlled, so that on some occasions that control can be lost. That seems to require meta-management processes that can monitor, and evaluate and control management processes.

And I don't believe our architecture is the only one or the right one, though it does seem to make sense of such processes. There may be other architectures which could do just as good a job, or a better one.

A.9 Could current network server architectures suffice?

??: I have a comment, and I think it's worth looking at it just a second:

If you go back to the babies problem, I am pretty convinced after that comment that you could probably formulate that as a network server model for real-time systems. There's nothing in that that's particularly unusual: you have got resource limitations, you have got reconfigurable networks. You might have a very complicated system, but you are basically serving customers.

And if you allow this system to run in an experimental situation for some period of time, you can probably collect enough statistics to define your heuristics or optimizations, or your local optimizations, that will give you a static state behaviour. Well, if you don't have static state behaviours, then you either have a situation where everyone is happy, or where someone is frantic or desperate, if you want an emotional content. But there's not anything in your example which I find to violate the kinds of formal models that people already use for those kinds of situations. In other words, I am not convinced that this has to be approached as an AI problem, first.

Sloman: Well, there are networks where there is a lot of distributed control about routing. We are not talking about that. We are talking about something like, say, a file-server, which is serving lots of different computers.

So, the next question is, how many different sources of motivation does your network server have.

??: Well, all those things that can request output, each of which is like a baby.

Sloman: Right. So, we have a class of requests for some services, which can arrive asynchronously and have varying importance and urgency.

??: Yes, it may have finite buffers, which, if they overflow, it's equivalent to death or injury.

??: I think maybe there is a better problem domain which would express your ideas better, because –

Sloman: There may be millions of analogous problem domains. But the question I was going to ask was, where does the need to plan come into the network server?

??: It depends on whether you believe the problem has a solution, in which case you try to perform an optimization, or whether you simply do the best you can. It seems to me that there are just numerical algorithms, or statistical algorithms, to maximize the likelihood of successful operation.

Sloman: What I am claiming is that a human being, in this situation, will find things difficult. And that's partly what I want to explain.

I am not trying to find an optimal engineering solution to the problem. If I were I would be looking for a different sort of architecture with different sorts of mechanisms.

BLW: Right. You wouldn't consider an architecture with a planner, and then a planner and a meta-management.

You would collect statistics, trying out a few ways of doing the plans, figure out the best strategy, on the basis of the distribution of baby arrival times, and distribution of baby death times.

Sloman: Yes, and in fact, a real nursemaid might actually eventually learn some of that, and use that to control fast heuristics for deciding what to do. Learning of that kind is something that we eventually need to build into our system.

Such mechanisms, designed by evolution and tuned by adaptive processes in individuals, could play an important role in the automatic processes that we share with other animals. But they are not flexible enough when you need to deal with circumstances that can change too much for statistical adaptive processes to cope, e.g. because you need to be able to construct new plans, make complex inferences, form complex new systems of concepts.

The niche to which a human architecture is fitted is far more complex than a typical network server's niche, at present anyway.

BLW: But is it so complicated that you need this kind of system. Or is there a point at which you can just have this kind of scheduling system, which is optimized under these conditions do this.

Sloman: No matter what the external problem is, there will always be an architecture that solves this problem quite differently from the way we do, by having very much faster processing internally, and different kinds of memory, or whatever. So, to that extent, I am not claiming uniqueness.

??: Yes, I meant, the distinction is not to solve the nurse-maid problem, to create the optimal nurse-maid. The goal is not to make that particular nurse-maid work well, but to make the nurse-maid that's .... (tape not clear)

Sloman: And if it turns out to be too easy to do it in the nursery domain, I have indicated that there are lots of ways of making the environment more complicated, and more like real human life.

However, I regard as obviously false the claim, fashionable in some quarters, that human-like intelligent agents never have to do any explicit plan construction and problem solving prior to performing actions. For example before coming to this workshop I had to explore alternative dates and modes of travel and create a plan, including deciding whether to go to the airport by train, coach or car. That took time and interfered with other activities, such as planning lectures, etc.

So the kind of niche that defines our research problem *requires* the minder to be able to perform planning tasks that take time, even if a competent engineer could find a solution for *this* simple domain that made use of pre-stored plans and condition-action rules.

A.10 How should an artificial domain be evaluated for realism?

BLW: Well. You will have to be able to evaluate whether this is or isn't comparable to humans. You need some kind of metric given this artificial world, saying what is or isn't, comparable, and saying what a human would do under these circumstances.

And given it's not an optimization problem, where you can say you have or haven't got an optimal solution, what sort of things are you using to develop intuitions for saying whether the decisions you have made are the ones that a person would make?

Sloman: I don't have any direct answer to that. First of all, there is no such thing as *the* decision a person will make, because people are all different. So, there might be classes of decisions, and we might have to find a framework within which one –

BLW: But you have to be able to say that this time, it didn't make the right decision ...... and say: something is missing from the architecture.

Sloman: Yes, and you might say that the reason it did not make the right decision is the same as the reason why a human being in the same situation might not make the right decision.

So, it made a 'right' decision in terms of the time pressure.

BLW: Allowing for that, you have to be able to characterize what's happening.

Sloman: Sorry, I started by making the negative point about the difficulty of making the comparison with humans, which I shouldn't have done. It's a distraction. And I'll now try to give the positive answer.

First, at a lower level, we have direct ways of evaluating the system in the way you are asking for e.g. by asking human beings to drive the simulation and comparing their behaviour with that produced by a software system.

We also have indirect ways, which will take a long time.

One of these indirect ways is by looking to see how well this framework generates new explanations for things that psychologists and psychiatrists and social workers and teachers are concerned about.

So, for example, recently, with two of my students, I produced the previously mentioned paper on emotions like grief (Wright et al. 1996). What triggered this was my reading an autobiographical report of a particular experience of intense and long lasting grief, written by someone whose much loved friend had died following a long illness.

In our paper we made a first shot at trying to show how the phenomena might come out of something like our architecture. In fact, the process of doing that forced us to pay attention to some things missing from the architecture, including support for experiences of pleasure and pain, and certain kinds of self-consciousness and self-evaluation, to define what it meant to be in control of one's own thought processes.

This led to further developments in the design, though not yet to implementation. Thus comparison of a model with detailed phenomenological reports is one sort of evaluation.

Another is what happens when I show a paper like that to experts from other fields, e.g. clinical psychologists. To my pleasure and surprise, some of them are already saying, that the paper is relevant and helpful to them. In fact, when it was read by the editor of a journal on Philosophy, Psychology and psychiatry, he immediately wanted to publish it. That does not prove that our ideas are correct. It's only a partial initial indication that we may be making progress in a useful direction.

I regard this as a kind of indirect test, which doesn't say our ideas are right. It just says that we seem to have some useful ideas that other people haven't got. But it may turn out that something else is even better. And that's how I think all science goes. You never can prove anything conclusively correct or incorrect: you can only gradually get closer and closer to accurate theories.

A.11 Testing by running a simulation

BLW: So, you point to some event that's happened in a simulation. And you say: why did this happen? You may find as well as getting behaviour that seems to correlate with what people do, you also observe behaviours that don't correlate with what people do. And you may find an explanation in terms of some features of the architecture. Will you then go to a psychiatrist and say, do you ever come up with an example of this?

Sloman: We might. And that would be interesting. However, as I explained previously I think the problems we are addressing are enormously complex and certainly will not be solved in my lifetime. Moreover our current thinking may be limited by our current understanding of architectures mechanisms and formalisms.

So for some time I don't expect that the main form of test will come by studying the actual behaviour of a working implementation of our ideas. Rather, an earlier phase, the mere task of planning the implementation is making us discover gaps in our thinking, making us understand the problem better, sending us back to the drawing board to extend either the requirements specification (the niche description) or some aspect of the design.

For that we don't need implementations yet, though we do need to be trying to produce them.

However, implementations also have a cosmetic value. People will pay more attention to our ideas if we can show them something working. It proves it's not all empty hand-waving if it can drive development of a real implementation, unlike many so-called theories. And we also hope eventually to produce a nice teaching tool for psychology students and others.

BLW: But don't you get something out of running the implementation?

Sloman: I will get nothing out of running it, I think. That's my guess.

??: ....

Sloman: Well, what I really want to do, is explore and understand design space. When I said I get nothing out of running it, that was a bit extreme – certainly it has some impact on our thinking.

And we may well get surprising behaviour. We may have to say, oops, there is something we haven't thought about, which we have to get back to and try to understand.

But equally, I personally don't want to do a long line of experiments comparing an implementation with human behaviour when I know in advance that there will be large discrepancies because our implementations in the foreseeable future will be very much oversimplified, and also because people are so variable.

I regard it as more important to do a deep analysis of the problem, asking what effects differences in designs will have in different situations. E.g. how will they affect speed or correctness or precision of performance? I always want to move up to a higher level of understanding of what's going on, and then perhaps do some tests to see whether the understanding is right. But mainly that test will consist in seeing whether the ideas are implementable.

But I don't simply want to run a lot of tests to see what the program does if, instead, I can formulate theoretical predictions of what the architecture is and is not capable of. Of course, sometimes we get useful surprises from running the programs. But that in itself will be of no value, unless it can be related to a theoretical explanation of what's going on. When you have understood something general, as a result of that surprise, that will be of value.

A.12 Can the architecture model different personalities

BP: A different question: Given the architecture you propose, how do you start to model different personalities? Will they be just variations of the different components, like the way you said the filter, or the change of management strategy, or would it be something that would be completely different?

Sloman: In a real biological population, which is to have a social structure and a lot of cooperation, the same general system has to generate some differences within individuals to start with, e.g. to support divergence of function in the society.

Some of these differences will then be amplified by differences in individual experiences, e.g. either growing up in Vienna or growing up in an African jungle, or whatever.

Although there may be minor variations within the architecture, I would expect many different kinds of personalities to be accommodated within the same general architecture, e.g. different sorts of motive generators, different strategies for evaluating and comparing motivators, different ways of assigning priorities and thresholds, different meta-management strategies, and also many differences within the pre-attentive automatic part of the architecture, about which I have said very little because that's not the main focus of my research.

In the long run, we need to explore types of genetic differences in human beings and see whether we could find ways of implementing them. That raises many interesting questions: What is it that makes some of us want to be surgeons, and others want to be philosophers, while some people are happy to be airline pilots or bus drivers? There clearly are different kinds of life-styles and life preferences.

I do not claim that this is all genetically determined. It depends also on the extent to which individuals absorb information and strategies from the environment, e.g. how they generate new perceptual categories and new motivators, or motivator-generators, or motivator-comparators through the process of growing up in a particular culture. Some of these may be regarded as a change of architecture, e.g. acquisition of new types of abilities and new links between components. Remember that personality is not one thing but a large collection of information and control structures distributed throughout the system.

That's really long term research.

I suspect we can explain within our general *sort* of architecture some very interesting individual variations in terms of the kinds of ways different things are valued, and how different agents react to the same situation.

There won't be enough variety in our little toy domain to support all of that. We would need a much richer scenario to support individual variations of the sort that humans have, including different environments in which they develop. There may be a large set of detailed problems that each individual has to solve because of the structure of the environment and which produce long term changes affecting future processing and behaviour.

So, a full study of personality would require us to investigate the whole range of different ways in which individuals can vary, both genetically and in their development, despite sharing a common generic architecture, at least at birth.

Whether and how we will ever be able to implement them, I don't know. Only a tiny subset will be done in my life-time, that's for sure.

References

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

L.P. Beaudoin and A. Sloman, A study of motive processing and attention, in *Prospects for Artificial Intelligence,* eds A.Sloman and D.Hogg and G.Humphreys and D. Partridge and A. Ramsay, 229–238, IOS Press, Amsterdam, 1993,

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

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

Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

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

Marr, D. *Vision* Freeman, 1982

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

A. Ortony and G.L. Clore and A. Collins, *The Cognitive Structure of the Emotions,* Cambridge University Press, New York, 1988,

Pryor, L., & Collins, G. (1992). Reference features as guides to reasoning about opportunities. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Bloomington, Lawrence Erlbaum Associates.

Ryle, G. (1949) *The Concept of Mind,* Hutchinson.

H. A. Simon, Motivational and Emotional Controls of Cognition, 1967, Reprinted in *Models of Thought,* Yale University Press, 29–38, 1979

A. Sloman, *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind,* Harvester Press (and Humanities Press), 1978, Hassocks, Sussex,

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

A. Sloman, Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence, *Proc 2nd Int Joint Conf. on AI,* 1971, London, Reprinted in *Artificial Intelligence,* 1971

A. Sloman, 1985, What enables a machine to understand?, *Proc 9th Int Joint Conf on AI* Los Angeles, 995–1001

A. Sloman, Motives Mechanisms and Emotions', *Cognition and Emotion* 1987, 1, 3, 217–234, Reprinted in M.A.Boden (ed), *The Philosophy of Artificial Intelligence,* OUP, 1990,

A. Sloman, On designing a visual system (Towards a Gibsonian computational model of vision), *Journal of Experimental and Theoretical AI,* 1989, 1, 4, 289–337

A. Sloman, Notes on consciousness, *AISB Quarterly,* 1990, 72, 8–14,

A. Sloman, Prolegomena to a theory of communication and affect, in A. Ortony and J. Slack and O. Stock, (eds), *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues,* Springer, 1992, 229–260, Heidelberg, Germany

A. Sloman, Prospects for AI as the general science of intelligence, in (eds) A.Sloman, D.Hogg, G.Humphreys, D. Partridge and A. Ramsay, *Prospects for Artificial Intelligence,* IOS Press, 1993, 1–10, Amsterdam

A. Sloman, The mind as a control system, (eds) C. Hookway and D. Peterson, *Philosophy and the Cognitive Sciences,* Cambridge University Press, 1993, 69–110

A. Sloman (1994a) Explorations in Design Space, *Proc 11th European Conference on AI,* 1994, Amsterdam,

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

A. Sloman, Exploring design space and niche space, *Proc. 5th Scandinavian Conf. on AI,* Trondheim, 1995, IOS Press, Amsterdam

Sloman, A and Poli R, (1996) SIM_AGENT: A toolkit for exploring agent designs, in *Intelligent Agents Vol II (ATAL-95)*, Eds. Mike Wooldridge, Joerg Mueller, Milind Tambe, Springer-Verlag pp 392–407 Presented at *ATAL-95, Workshop on Agent Theories, Architectures, and Languages,* IJCAI-95, Montreal, August 1995.
(Also Cognitive Science technical report: CSRP-95-4)

I. Wright, A Summary of the Attention and Affect Project, 1994, Available at URL:
ftp://ftp.cs.bham.ac.uk/pub/dist/papers/cog_affect in the file Ian.Wright_Project_Summary.ps.Z

Wright, I.P, Sloman, A, & Beaudoin L.P (to appear, 1996) Towards a Design-Based Analysis of Emotional Episodes, To appear, with commentaries, in *Philosophy Psychiatry and Psychology*