Designing Human-Like Minds

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

Abstract

Under what conditions are "higher level" mental concepts which are applicable to human beings also applicable to artificial agents? Our conjecture is that our mental concepts (e.g. "belief", "desire", "intention", "experience", "mood", "emotion", etc.) are grounded in implicit assumptions about an underlying information processing architecture. At this level mechanisms operate on information structures with semantic content, but there is no presumption of rationality. Thus we don't need to assume Newell's knowledge-level, nor Dennett's "intentional stance." The actual architecture will clearly be richer than that naively presupposed by common sense. We outline a three tiered architecture: with reactive, deliberative and reflective layers, and corresponding layers in perceptual and action subsystems, and discuss some implications.

1 Introduction

We often describe other animals using mentalistic language. Some people assume that eventually software and/or hardware artificial agents will also merit such descriptions, though many disagree. McCarthy [4, 5] gives reasons why we shall need to describe intelligent robots in mentalistic terms, and why such a robot will need some degree of self consciousness, and he has made suggestions regarding the notation that we and the robot might use to describe its states. This paper extends that work by focusing on the underlying "high level" architectures required to justify ascriptions of mentality, ignoring, for now, questions about formalisms used within the architectures. We assume that there will be many different formalisms, serving different purposes [9, 14, 17].

Which particular mentalistic concepts are applicable to a system will depend on the architecture of that system. An architecture provides a basis for a family of interrelated concepts, namely concepts describing the states and processes able to occur in the architecture. It is conjectured that the normal adult human architecture involves three main layers, each supporting different sorts of mental concepts. The first layer is also the oldest in evolutionary terms, and probably shared with most other animals, and is entirely reactive. This is the focus of many evolutionary alife experiments, and much explicit design work.

The layer second is deliberative. This is newer and probably found in far fewer animals. The emergence of significant deliberative capabilities in alife experiments may require considerable increases in computer power, as the mechanisms are very different from those required for reactive systems (as we'll see below).

The third is a reflective layer (previously labelled "meta-management") which is much newer and probably much rarer. The evolutionary step from deliberative to reflective mechanisms is probably much smaller than that from reactive to deliberative. That is because deliberative mechanisms (and to some extent reactive mechanisms) can be re-used in reflective mechanisms.

These layers, about which more is said below, support different sorts of mental processes, including motivational and emotional processes, a fact that has not been noticed among emotion theorists, leading to a plethora of different definitions of "emotion" and unrelated theories of emotion, which appear to contradict one another, but are actually talking about different things. Our diagnosis is that those who stress emotions based on the limbic system and observable in rats and most other animals are studying effects of the reactive layer. Those who stress emotions such as apprehension, disappointment and relief, related to phases in the execution of plans, are studying effects of the deliberative layer. By contrast poets, novelists and those who study emotions involving loss of control of thought processes (e.g. our work on grief [18]), are studying processes involving the reflective, or meta-management layer.

1.1 An example: self-control and emotions

An example will illustrate how mentalistic descriptions can depend on an architecture. We talk about humans sometimes losing control of themselves, for instance in certain emotional states. This presupposes the possibility of transitions between being in control and losing self control. That in turn depends on the existence of an architecture that supports certain kinds of self monitoring, self evaluation, and self modification including decisions about what to attend to or think about.

For systems lacking the architectural underpinnings, certain descriptions of mental states and processes (e.g. "emotional", "restrained", "resisting temptation") may be inapplicable.

Whether other animals have architectures that can support these descriptions is not clear. Neither is it clear what sorts of architectures in software agents will make such states and processes possible. We have some tentative suggestions outlined below.

Our claim is *not* that such architectural features are required to produce human-like behaviour. Any form of behaviour can be produced by an indefinite variety of significantly different architectures. In particular, any finite sequence of behaviours produced by our three layered architecture could, *in principle*, be produced by a purely reactive system. However, the latter would have to be far more complex (having far more pre-designed behaviours) and would take far longer to evolve, since all the different forms of behaviour would have to be separately selected and encoded in genetic structures, instead of being created when required at "run time" by a deliberative mechanism. From this standpoint much of the current posturing about oppositions between reactive and deliberative approaches is just ill-informed and sometimes rather childish especially since the concepts involved in reactive and situated systems are very old in cognitive science and AI (e.g. [3, 8]) even if the terminology is new.

1.2 A comparison: the architecture of matter

The relationship between mental concepts and the underlying architecture can be compared with the way in which a new theory of the architecture of matter generated the table of possible elements: the periodic table. Within the framework of the atomic theory of matter developed during the last two centuries, it became possible to see which previous concepts of "kinds of stuff" were suited to describing the physical world and which ones needed to be refined or rejected. The new architecture also revealed the need for a host of concepts for kinds of physical matter that had not previously been thought of, e.g. elements whose possibility was first revealed by the periodic table.

Similarly a good theory of the architecture of a type of agent is likely to show the need for revisions and extensions of our existing theory of types of states in such agents. In particular, by examining deviations from a normal architecture, including possible genetic malformations and deviations caused by disease or damage, we can extend our class of concepts for describing mental states and processes. This could provide a far better way of classifying types of mental disorder, as well as preparing us for some of the bizarre possibilities to be expected in artificial agents whose architectures differ in various ways from ours.

2 Mentalistic concepts applicable to artificial agents

It is often convenient to describe a machine as "choosing", "exploring", "deciding", "inferring", etc. The states and processes referred to are *intentional*, since they have semantic contents. In some cases it may be useful also to describe such systems as "believing", "wanting", "preferring", "enjoying", "disliking", "frightened", "angry", "relieved", "delighted".

If applying such mentalistic concepts to people implicitly assumes a certain sort of high level information processing architecture, then similar architectural requirements will need to be satisfied by artificial agents if applying mentalistic terms to them is not to be misleading, like the over-enthusiastic use of words like "goal" and "plan" in some early AI publications, criticised by McDermott [6]. (Similar criticisms can be made of the words like "learn" and "discover" applied to neural nets. There is no monopoly on hyperbole.) Of course, there may be interesting intermediate cases based on different architectures.

All this assumes that purely behavioural definitions of mentalistic concepts (in terms of relationships between externally observable inputs and outputs) cannot adequately define these concepts. This anti-behaviourist stance has a long history (e.g. [1]) and will not be defended here.

It is not claimed here that humans acquire their assumptions about mentalistic architectures either by introspection or by induction from observations of the behaviour of others. Rather it is conjectured that just as animals may be born with innate predispositions to interpret their physical environment within a certain class of ontologies, similarly social animals may be born with innate predispositions to postulate "internal architectures" that can be used to categorise, explain and predict the behaviour of others (and themselves). In both cases the details may be shaped by physical and cultural influences during individual development.

From this viewpoint, evolution, or rather co-evolution, not philosophical argument, solved the mind-body problem for us. However, this is a complex topic which will not be pursued further here.

3 Why use mentalistic language?

We shall need mentalistic descriptions for artificial agents (a) because of marketing requirements, (b) because such descriptions will be irresistible and (c) because no other vocabulary will be as useful for describing, explaining, predicting capabilities and behaviour. ((c) provides part of the explanation for (b).)

E.g. descriptions in terms of physical processes, or the programming language level data-structures and algorithms will not be useful for those who have to interact with the agents, however useful they are for developers and maintainers. This is analogous to the fact that interacting with people is difficult if the only way you can think about them is in terms of their internal physiological states.

So, instead of trying to avoid the use of mentalistic language, which will be self-defeating, we need a disciplined approach to its use. This can come by basing mentalistic concepts on architectural concepts: i.e. we use the 'design stance'.

Unlike Dennett and Newell ...

This differs from the approach of Dennett [2] who recommends the "intentional stance" in describing sophisticated robots, as well as human beings. This stance presupposes that the agents being described are rational: otherwise their behaviour provides no basis for inferring beliefs, desires, intentions, etc.

Our stance also differs from the approach of Newell [7] who recommends the use of the "knowledge level", which also presupposes rationality.

... We use an "information level" design stance

Our claim is that mentality is concerned with an "information level" architecture, close to the requirements often specified by software engineers, for instance in designing an office information system or a factory control system [12]. Such systems involve components that acquire, manipulate and use information, including information about objects in the environment. This is a version of the design level of description, which lies between physical levels (including physical design levels) of description and intentional descriptions that always refer to the whole agent.

The "holistic" intentional stance permits only talk about what *the whole* agent believes, desires, intends, etc. Information level design descriptions also allow us to talk about various semantically rich *internal* information stores, motive databases, and state transitions that are possible for internal information items (e.g. being generated, evaluated, adopted, rejected, stored for future consideration, interrupted, suspended, reactivated, modified, destroyed, matched against other items, etc.)

4 Rationality is not a requirement for mentality

The mechanisms in such an architecture need be neither rational nor irrational: even though they acquire information, evaluate it, use it, store it, etc. [12]. Some of the processes are simply *automatic*. The system could be non-rational because only a small subset of its behaviour is based on explicit, disciplined, evaluation of alternatives in the light of a consistent set of beliefs and a totally or partially ordered set of desires and preferences. E.g. much of the behaviour even at a high level may be *habitual*. In addition there are phenomena such as carelessness, impulsiveness, laziness, etc. Architectural requirements for carelessness are described below.

The claim being made here is that lack of rationality does not prevent processes and mechanisms being concerned with semantic information (including internal references: such as one internal structure that is used by the machine to describe the relationship between two other structures, for instance a history of changes in plans which may be useful in preventing looping and other wasted actions during planning). In short *intentionality* does not require *rationality*.

There is no commitment at this stage regarding the *form* used to encode or express information. It may include logical databases, procedures, image structures, neural nets or in limiting cases physical representations, such as curvature of a bimetallic strip representing temperature. (For more on this see [14, 16, 17].)

At this level we can begin to explain what mental states are in terms of the information processing and control functions of the architecture. These functions include having and using information *about* things. E.g. an operating system has and uses information *about* the processes it is running. Thus semantic content is already present, without full-blown intentionality or rationality.

By describing a variety of functions using the "design stance" at the information level, and showing how they implement mental states and processes, we provide a richer and deeper explanatory framework than the intentional stance.

5 Emergent states and processes

Not all states require specific supporting mechanisms in the architecture. A computing system that is "overloaded" does not have an "overloading" mechanism. Rather that is a feature of the interaction of many different mechanisms all of which have functions other than producing overload. Similarly with many mental states, e.g. feeling humiliated. There need not be a special humiliation mechanism.

If the system which combines reactive and deliberative mechanisms also has the reflective (meta-management) ability to monitor, evaluate and to some extent control its own states and processes then a new variety of descriptions becomes applicable, including new forms of self control, learning of concepts for self-description, etc.

In particular, the phenomena often described by philosophers and others as involving "qualia" may be explained in terms of high level control mechanisms that can attend to many internal states and processes including internal intermediate structures produced during the processing of sensory information (visual qualia). Likewise a robot able inspect its visual qualia, like a human visiting an oculist, may be able to contribute to diagnosis of malfunctions by describing its qualia ("horizontal lines look blurred").

The objects of such self-monitoring processes may be virtual machine states rather than internal physical or physiological states. Software agents able to inform us (or other artificial agents) about their own internal states and processes may need similar architectural underpinnings for qualia.

This need be no different from the mechanisms underpinning a child's ability to describe the location and quality of its pain to its mother, or an artist's ability to depict how things look (as opposed to how they are), or a patient's ability to tell an oculist about changes in the way things look as different lenses are tried. In the same way, robots of the future possessing a suitable architecture may be able to report many features of their experience as part of the process of detecting and diagnosing malfunctions.

Another example of the relationship between architectural underpinnings and mentalistic description follows.

6 Example: What is required for carelessness?

Describing X as "working carelessly" implies that
(a) X had certain capabilities relevant to the task in hand,
(b) X had the ability to check and detect the need to deploy those capabilities,
(c) the actual task required these abilities to be deployed (e.g. some danger threshold was exceeded, which could have been detected, whereupon remedial action would have been taken),
(d) something was lacking in the exercise of these capabilities on this occasion so that some undesirable consequence ensued or nearly ensued.

X's carelessness could have several forms:

- X forgets the relevance of some of the checks (a memory failure),
- X does not focus attention on the data that could indicate the need for remedial action (an attention failure),
- X uses some shortcut algorithm that works in some situations and was wrongly judged appropriate here (a selection error),
- X does not process the data in sufficient depth because of a misjudgement about the depth required (a strategy failure),
- X failed to set up the conditions (e.g. turning on a monitor) that would enable the problem to catch his attention (a management failure).

This illustrates how familiar mentalistic descriptions can be based on architectural concepts. Replacing the above characterisations in terms of observable features of a careless person, without referring to internal states and mechanisms, would require a very much larger description of a wide range of cases, and would inevitably lack the full generality of the "modular" analysis in terms of underlying mechanism. Similar remarks can be made about a very wide range of mentalistic concepts.

The presuppositions for "working carefully" are similar to those for working carelessly. Something that is incapable of being careless cannot be careful.

Our claim is that when people use mentalistic language to describe themselves or other humans they implicitly presuppose that there are various coexisting interacting subsystems with different functional roles, for instance, perceptual subsystems, various types of memory, various skill stores, motivational mechanisms, various problem solving capabilities.

There is no reason why we should not transfer these predicates to artificial agents, if they have appropriate architectures. Some systems which lack the architectures may still produce behaviour suggesting similar mental states to observers, especially as the human brain seems strongly predisposed to attribute mentality to mobile objects whose behaviour is not easily explained by observable physical forces. But where the underlying architectures are different, there may be surprises in some contexts.

However, it remains to be seen whether there are architectures that can produce human-like behaviour in a wide range of cases without involving the distinct kinds of structures outlined here. It is an empirical question whether this will turn out, in the long run, to be a good theory about how human minds actually work at some level of description. This could be true even if, as some suppose, the full complexity of any individual human mind will forever defeat any attempt at human understanding or replication. We may be equally unable to grasp or model the full complexity of an individual oak tree or thundercloud: but that does not mean the important general principles they instantiate are beyond our comprehension.

7 One way to make progress

There are many different, but important and valid, approaches to the study of human-like systems, and it is foolish to be prescriptive and suggest there is only one way. The only claim being made here is that among the *many* potentially useful approaches to the study of autonomous agents, whether natural or artificial, designed or evolved, is the approach based on attempting to devise a more accurate and explicit theory of the variety of types of architecture capable of producing the sorts of capabilities we wish to explain, model or replicate. We can then use such architectures as frameworks for generating families of descriptive concepts applicable to different sorts of humans (including infants and people with various kinds of brain damage) and different sorts of animals and artificial agents. Layered architectures of the sorts described here are offered simply as one important class for investigation. There may be many other classes of architectures all requiring detailed study.

The investigation of classes of such architectures *design space* and the kinds of requirements they satisfy *niche space* will require detailed multi-disciplinary collaboration. It is not a task that we can expect to complete in the foreseeable future, if ever. But that does not mean that progress is impossible. One way to make progress is to study small regions of niche space and design space. One such region, though obviously not the only one worth study, includes a neighbourhood involving architectures characterised as follows.

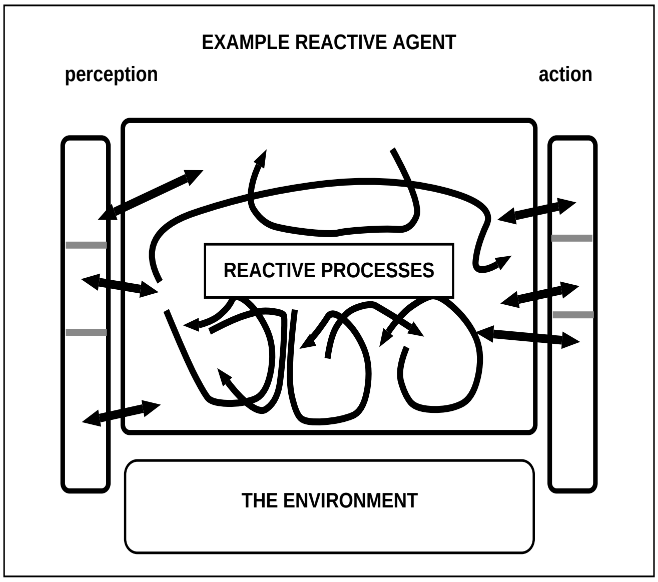
To recapitulate, the conjecture, which will now be elaborated, is that evolvable, physically implementable, human-like agents could be implemented in nature and possibly also in artificial systems with an architecture with at least three coexisting, interacting, concurrently active layers (see figures below):
• A very old reactive layer, found in various forms in all animals, including insects).
• A more recently evolved deliberative layer, found in varying degrees of sophistication in some other animals (e.g. cats, monkeys).
• An even more recent meta-management (reflective) layer providing self-monitoring and self-control, perhaps found in simple forms only in other primates. (Probably only in a very rudimentary form in very young children – i.e. the architecture is not static but develops).

8 Reactive agents

We now turn to a more detailed (but still incomplete) characterisation of the three layers.

In a purely reactive agent:

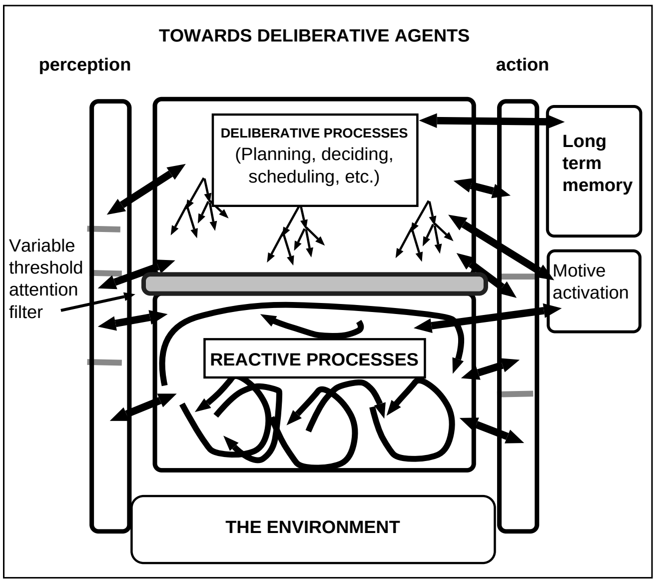
• Mechanisms and space are dedicated to specific tasks
• There is no construction of new plans or structural descriptions
• There is no explicit evaluation of alternative structures
• Many of the processes may be continuous rather than digital, e.g. analog feedback control circuits
• Parallelism and dedicated hardware can achieve high



speeds
• Conflicts may be handled by relatively simple mechanisms like vector addition or winner-takes-all networks.
• Some learning is possible, but only by modifying relative weights or association strengths between fixed categories of events: e.g. tunable control loops, change of weights by reinforcement learning
• There may be some hierarchical control structures, and some of these may benefit from hierarchic processes in perceptual and action sub-systems, e.g. if some of the higher level reactive behaviours are triggered by the output of higher level perceptual mechanisms (e.g. producing more abstract classifications)
• The agent can survive even if it has only genetically determined behaviours, provided that the environment does not present many problems for which the genetically determined solutions fail
• A reactive agent may be unable to cope in a situation which requires new plan structures, i.e. new combinations of actions which cannot be generated simply by constantly reacting in the current situation, but this lack of flexibility may not matter if such situations are rare, and individuals are cheap and expendable (e.g. insects are the most successful type of animal, in terms of proportion of biomass?).

9 Combining reactive and deliberative layers

What sort of evolutionary process could lead from purely reactive systems to systems combining reactive and deliberative mechanisms is not clear. It may be that an intermediate step would be the provision of a long term associative memory, which might, for example, be



useful both for modifying some of the reactive behaviours and also for learning useful geographical routes. Another intermediate step might be mechanisms which instead of directly generating actions to be performed instead generate goals to be achieved possibly by a succession of actions. This can certainly be part of a reactive mechanism, where one sort of internal reaction is the generation of a goal state, which then provides part of the context controlling other reactive behaviours. Later on such goal descriptions could drive processes which create *hypothetical* action sequences to be evaluated prior to actual execution. This requires some sort of temporary workspace in which such plans can be created and compared. Moreover, the implementation of a deliberative mechanism could use reactive mechanisms which operate not on the environment but on internal structures. None of this specifies in any detail how deliberative mechanisms could evolve. However I hope it is clear that although it may require some discontinuities in the evolutionary process (which are in any case required for Darwinian evolution) there is nothing essentially mysterious or obviously incapable of being part of the evolutionary process.

These are all topics for further research. For now, however, the following points should help to clarify some of the features of a hybrid system combining reactive and deliberative layers.

In a deliberative mechanism

• The key feature is that new plans may be constructed, composed of novel combinations of previously existing action steps (possibly also including conditional branches and loops, etc.).
• The process is deliberative in that newly created options can be explicitly evaluated before selection or execution: some plans will be created, evaluated and discarded, requiring alternatives to be created.
• The selection of a new step to add to an incomplete plan requires the use of a long term memory in which (a) the context of that incomplete plan can be used to select possible further steps and (b) the likely consequences of those steps are stored, on the basis of which the step can be assessed as good or bad.
• Since such plans are composed of collections of steps the process of construction is inherently discrete, not continuous. (It may be that there are also some plan modification processes involving continuous deformations of structures: this is one of many open research questions.)
• In humans (and some other animals?), learnt strategies created in the deliberative layer can somehow be transferred to the reactive layer (which requires spare capacity in the reactive mechanism) – examples are learning to drive a car, learning to sightread music, learning many athletic skills, etc.
• Because re-usable mechanisms and space are dynamically allocated to alternative plans or plan fragments, many of the processes are inherently serial, and therefore resource limited.
• Access to a content addressable long term memory may also be inherently serial, even though highly parallel mechanisms are used to implement the memory. For instance if different questions are put to the memory concurrently cross-talk may prevent answers to either question being accurate.
• A possible efficiency reason for limiting the parallelism in the deliberative mechanism is that if N processes occur in parallel the task of credit and blame assignment in learning which subset of the processes produce which effects is of the order of 2^N. This problem is exacerbated by the need to learn about delayed effects.
• The need for integrated control, e.g. prevention of simultaneous decisions to move in two different directions, may also require some sort of serial top level management process.
• A problem for a resource limited deliberative system is that a fast-changing environment can cause too many interrupts, e.g. generation of too many unimportant goals,
• Filtering interrupts via dynamically varying thresholds may help but does not solve all problems if the filter mechanisms are also resource limited and capable of error [15, 18].
• Perceptual mechanisms may need to be able to cope with more abstract descriptions if the deliberative mechanism is to achieve maximum benefit in the environment, e.g. perceiving that one is under threat from a predator, noticing the possibility of creating a new relationship between objects which will facilitate achieving some goal (e.g. moving X under Y will make

it easier to reach Y).

• Similarly it may be useful if the action subsystems can accept more abstract, high level instructions and automatically generate the fine grained control processes for performing complex acts.

In the diagrams, the horizontal divisions in the perceptual and action sub-mechanisms are meant to suggest the existence of different levels of processing all happening concurrently. It is possible for some of the more abstract perceptual and action mechanisms to be genetically determined (e.g. supporting the ability of a new born deer to run with the herd) while others are learnt (e.g. a child learning to read, or play the piano).

10 The need for internal self-monitoring (meta-management)

Deliberative mechanisms may be implemented by using specialised reactive mechanisms, which react to internal events by performing internal tasks, and which can interpret explicit stored representations of rules and plans. It is fashionable to deny the need for such things, but so far no argument has shown that they are dispensable in general.

Deliberative mechanisms might use evolutionarily determined strategies for planning, problem solving, decision making, evaluating, etc. However, if such strategies are unchangeable by the individual they might be too rigid for certain contexts, e.g. in social animals where cultures and the "built" environment change faster than the genotype can. For such animals higher level control and modification of such deliberative strategies and rules may be desirable. Internal monitoring mechanisms may achieve this if they:

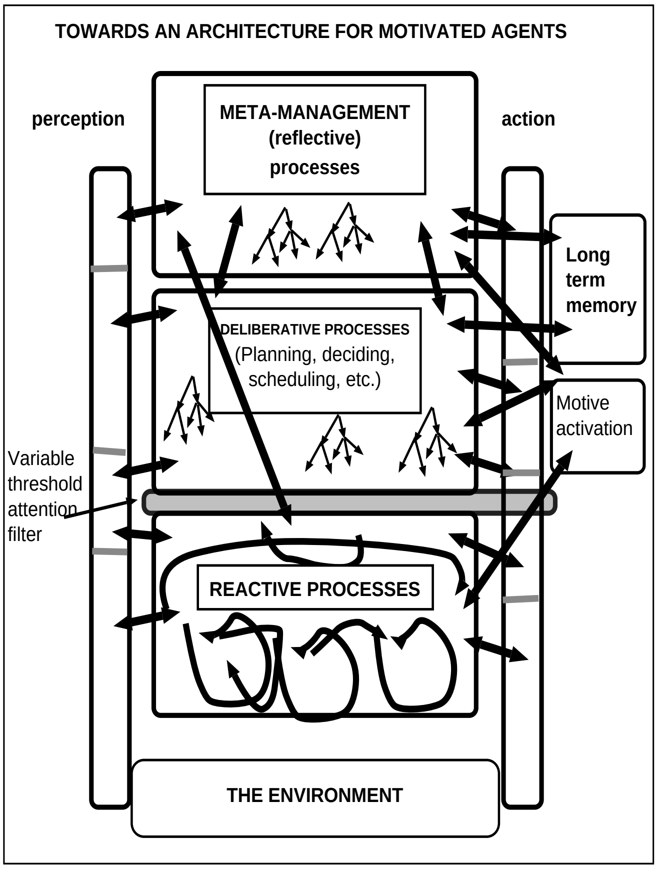
• Record events, problems, decisions taken by the deliberative mechanism, and notice patterns, such as that certain deliberative strategies work well only in certain conditions. This can lead to improved allocation of scarce deliberative resources.

• Allow exploration of new internal strategies, concepts, evaluation procedures, enabling discovery of new features, generalisations, categorisations.

• Allow diagnosis of injuries and illness by describing internal symptoms to others with more experience (e.g. a parent, an oculist).

• Evaluate high level thinking strategies, relative to high level long term generic objectives, or standards (e.g. deciding that one's deliberations are too self-centred, and thereby indirectly counter productive as a result of alienating others).

• Communicate more effectively with others, e.g. by using viewpoint-centred appearances to help direct attention ("A little to the left of where the hillside intersects the tree trunk"), or using drawings and paintings to communicate about how things look.



Meta-meta-management may not be needed if meta-management mechanisms are recursive (i.e. partly self-applicable)! However, humans seem to be limited in their ability to attend to their attending to their attending....

11 Generic functions of internal self-monitoring

"Meta-management" processes could include the following tasks.

• Reducing frequency of failure to achieve goals, by improving the order in which one thinks about sub-problems or the thinking strategies or algorithms used,

• Not allowing one goal to interfere with other goals, which requires improved strategies for detecting potential conflicts in advance,

• Not wasting time on problems that turn out not to be solvable,

• Not using a slow and resource-consuming strategy if a faster or more elegant method is available,

• Detecting possibilities for structure sharing.

There are probably many benefits that are restricted to social animals. For instance certain sorts of high level self-evaluations using criteria absorbed from a culture may be an important part of social control mechanisms using memes (whether for good or ill!).

12 Architectural layers and types of emotions

Among the many implications of having an architecture composed of the three layers is the possibility of very different sorts of mental states and processes, only some of which are shared with other animals that have simpler architectures [18]. In particular, different sorts of emotions seem to be associated with the different layers.

Many disagreements about the nature of emotions arise from a failure to grasp that there are different concepts of emotionality presupposing different architectural features, not all shared by all of the animals studied by emotion theorists.

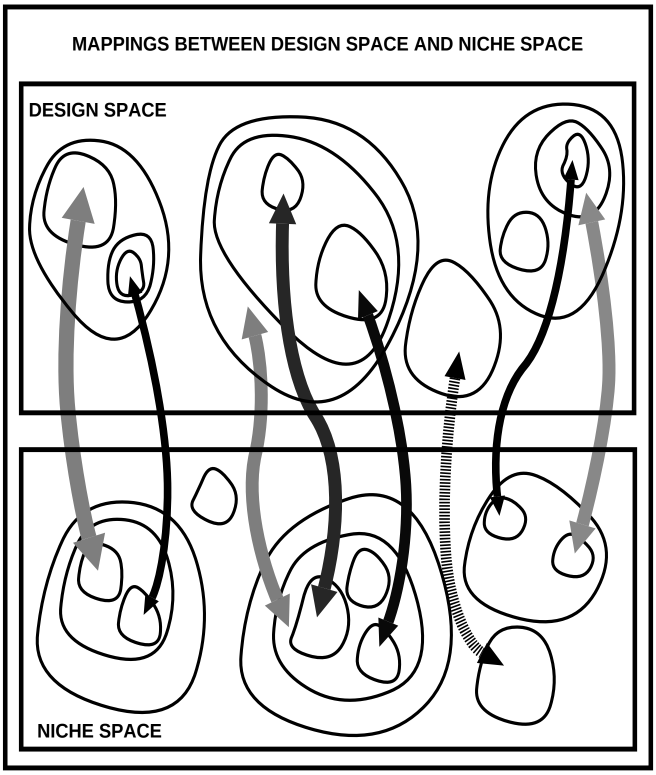
In particular, it is not always noticed that there are different sorts of *emotional* states and processes based on the different layers, e.g.:
(1) emotional states (like being startled, terrified, sexually stimulated) based on the old reactive layer shared with many other animals (i.e. using the limbic system?),
(2) emotional states (like being anxious, apprehensive, relieved, pleasantly surprised) which depend on the existence of the deliberative layer, in which plans can be created and executed, and threats, obstacles and opportunities can be detected and evaluated,
(3) emotional states (like feeling humiliated, infatuated, guilty, or full of excited anticipation) in which attempts to focus attention on urgent or important tasks can be difficult or impossible, because of processes in which the meta-management layer is frequently diverted by motives that get through the attention filter even though they have already been rejected (e.g. deciding not to go on thinking about how to have your revenge on the person who humiliated you, and then finding that you are thinking about it after all).

The second class of states depends on abilities that appear to be possessed by fewer animals than those that have reactive capabilities. The architectural underpinnings for the third class are probably even more rare: perhaps only a few primates have them. (Do rats lose control of thought processes?)

Within this framework we can dispose of much argumentation at cross-purposes, where people dispute about different sorts of things without a theoretical framework in which to discuss the differences.

13 There is no correct architecture

Different kinds of meta-management are likely to be found in different animals. Different architectures will be needed for different sorts of organisms or artificial agents. Even humans differ from one another. Architectures differ between between human children, adolescents, adults and senile adults. There may be culturally determined differences in architectures.



Similarly, naturally occurring alien intelligences and artificial human-like agents may turn out to have architectures that are not exactly like those of normal adult humans. Different architectures support different classes of mental states, so we shall need to be careful about assuming that existing forms of description, or even existing questions (What does it believe? What does it want? What is it trying to do? Why is it doing it?) are applicable to new sorts of agents.

Designers of synthetic agents need to be aware of the evolutionary pressures that led to these layers in human beings. Perhaps they are also required for certain classes of sophisticated artificial agents, whether robots or software agents.

In that case, there may be some unanticipated consequences of these design features [15].

Analysing these possibilities is hard. By developing a theory of a space of possible architectures [10, 11, 12, 13] we provide a framework for more precise specifications of alternative families of mentalistic concepts.

More specifically we need to explore relationships between "design space" and "niche space".

14 Design space and niche space

• A niche is a set of requirements, which may be capable of being satisfied in many different ways.
• A design is a set of specifications which may be capable of being instantiated in many different implementations.
• A physical environment does not uniquely determine

a niche: two animals in the same same geographical location may "inhabit" very different niches.

• Mappings between designs and niches are not unique: there are always trade-offs. Design A need not fit a niche better than design B in all respects.

• Designs need no designer, requirements no requirer.

One of the deep and difficult questions about these spaces concerns the types of trajectories that are possible subject to various constraints. One constraint is that the individual should adapt or modify itself to move from one region of design space to another, as, for instance humans do during individual development from infancy to adulthood. But some transitions are not possible for an individual (e.g. an acorn cannot possibly grow into a giraffe). However there are trajectories that are possible via evolution across generations that are not possible within individuals. Perhaps some trajectories are not possible via natural evolution but are possible via engineering processes.

Where species interact, the dynamics of such trajectories will generally be very complex and subtle, involving interactions between different niches and different designs, at different levels of abstraction.

These considerations help to put the Turing test in perspective. The test defines a tiny niche region of relatively little interest, except as a technical challenge. A small subset of design space will be relevant to it. From this perspective, its importance has been much inflated (though not by Turing himself).

15 More on the information level

Information level analysis presupposes that there are various information rich internal structures within the architecture. These need not be physically demarcated: they could be interacting structures in a virtual machine (as explained in [13].

The functional roles of such structures and substates are determined by such things as:

(a) where the information comes from,

(b) how it is stored,

(c) how it is processed or transformed before, during and after storage,

(d) whether it is preserved for a short or long time,

(e) how it can be accessed,

(f) which other components can access it,

(g) what they can do with the information,

(h) whether it actively generates new processes.

Notions of belief, imagining, reasoning, questioning, pondering, desiring, deciding, intending, having a mood, having an attitude, being emotional, etc. all presuppose diverse information stores with diverse syntactic forms, diverse mechanisms for operating on them, diverse contents and functional roles within the architecture.

However, it may turn out that for many architectures, including some found in nature and in artificial agents, our familiar modes of description may not be appropriate. For those we'll need to develop new systems of concepts and explanatory principles. (Can a goldfish long for its mother, and if not why not?)

These mental states do not presuppose rationality because many interactions between the components can produce irrational decisions or actions. For instance irrational impulses can be a product of an information processing architecture part of which is highly reactive.

16 Conclusion

A framework has been presented for multi-disciplinary investigation of many types of architecture of varying degrees of sophistication, with varying mixtures of information-processing capability; using AI, Alife, Biology, Neuroscience, Psychology, Psychiatry, Anthropology, Linguistics and Philosophy. In part this is based on an important level of analysis to which the design stance can be applied: the information processing level.

This is close to but different from Dennett's intentional stance and Newell's knowledge level, partly because it is concerned with lower level mechanisms (often virtual machines) for which considerations of rationality do not arise. Moreover, any general theory of agents should not focus on rationality as a central criterion of agency. It might rule out humans!

Even folk psychology allows for impulses, obsessions, addictions, memory lapses, various kinds of carelessness, temporary misjudgements of relative importance, and so on. Professional counsellors and therapists have additional ways of categorising mental states and processes without presupposing rationality (though which of them will survive creation of good theories about the underlying architecture is an open question).

People often need professional help, but the professionals don't always understand normal functioning, and therefore cannot account for deviations from normality, nor provide help reliably (except in the case of clearly defined physical and chemical abnormalities which can be remedied by drugs or surgery).

Similar possibilities arise for sufficiently sophisticated artificial agents. Artificial agents may also need therapy and counselling, for the same reasons as humans [15]. Existing human therapies may fail for the same reasons.

All this work, and especially the study of processes supported by different sorts of architectures, may force us to invent new concepts for describing some sorts of synthetic minds as well as providing better ways of talking about ourselves and other animals.

The ideas sketched here are still incomplete (what are pleasures, pains, aesthetic states?) and conjectural, but not wild conjectures: they have evolved from previous ideas through confrontation with design requirements and empirical facts from psychology, brain science and animal research. In the longer term, testing these ideas will require collaborative work in a range of disciplines including mathematics, computer science, software engineering, AI, Alife, brain science, clinical and developmental psychology, anthropology, ethology, evolutionary biology, etc. These different types of exploration should proceed in parallel, with people talking to and learning from one another instead of making silly dogmatic claims about there being only one way to make progress.

Acknowledgements and Notes
This work has been supported by the UK Joint Council Initiative, The Renaissance Trust, DRA Malvern, and the University of Birmingham. Much has been learnt from research students and staff at Birmingham, in the Schools of Computer Science and Psychology, especially Luc Beaudoin, Chris Complin, Darryl Davis, Glyn Humphreys, Brian Logan, Christian Paterson, Riccardo Poli, Tim Read, Ed Shing, Ian Wright. A toolkit developed jointly with Riccardo Poli and used for exploring a variety of types of agent architectures and doing evolutionary experiments is described in
http://www.cs.bham.ac.uk/~axs/cog_affect/sim_agent.html

Several papers developing these ideas are in the Cognition and Affect Project directory. See:
ftp://ftp.cs.bham.ac.uk/pub/groups/cog_affect
http://www.cs.bham.ac.uk/~axs/cog_affect

References

[1] Noam Chomksy. Review of skinner's *verbal behaviour*. *Language*, 35:26–58, 1959.

[2] D. C. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology.* MIT Press, Cambridge, MA, 1978.

[3] G.A.Miller, E.Galanter, and K.H.Pribram. *Plans and the Structure of Behaviour.* Holt, New York, 1960.

[4] J. McCarthy. Ascribing mental qualities to machines. In M. Ringle, editor, *Philosophical Perspectives in Artificial Intelligence*, pages 161–195. Humanities Press, Atlantic Highlands, NJ, 1979. (Also accessible at http://www-formal.stanford.edu/jmc/ascribing/ascribing.html).

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

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

[7] A. Newell. The knowledge level. *Artificial Intelligence*, 18(1):87–127, 1982.

[8] H. A. Simon. *The Sciences of the Artificial.* MIT Press, Cambridge, Mass, 1969. (Second edition 1981).

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

[10] A. Sloman. The mind as a control system. In C. Hookway and D. Peterson, editors, *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press, 1993.

[11] A. Sloman. Explorations in design space. In *Proceedings 11th European Conference on AI*, Amsterdam, 1994.

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

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

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

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

[16] Aaron Sloman. Actual possibilities. In Luigia Carlucci Aiello and Stuart C. Shapiro, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, pages 627–638. Morgan Kaufmann Publishers, 1996.

[17] Aaron Sloman. Towards a general theory of representations. In D.M.Peterson, editor, *Forms of representation: an interdisciplinary theme for cognitive science.* Intellect Books, Exeter, U.K., 1996. ISBN: 1-871516-34-X.

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