Damasio, Descartes, Alarms and Meta-management

Aaron Sloman
School of Computer Science
The University of Birmingham,
Birmingham, B15 2TT, UK

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

ABSTRACT

This paper discusses some of the requirements for the control architecture of an intelligent human-like agent with multiple independent dynamically changing motives in a dynamically changing only partly predictable world. The architecture proposed includes a combination of reactive, deliberative and meta-management mechanisms along with one or more global "alarm" systems. The engineering design requirements are discussed in relation our evolutionary history, evidence of brain function and recent theories of Damasio and others about the relationships between intelligence and emotions.

1. INTRODUCTION

Stan Franklin, the organiser of this symposium, wrote "Minds are the control structures of autonomous agents" [5, p 412]. The claim that minds are essentially concerned with control, echoing the seminal ideas of Norbert Wiener [16] is one with which I strongly concur though as argued in [11], we need to go far beyond the early idea of control systems with fixed architecture and changes only in numeric values.

The most obvious characteristic of our minds, which is perhaps least noted in philosophy,, is that they are *active*. This does not mean that the function of mind is constantly to be moving our arms and legs and other physical components. For humans, most of the activity is *internal*: noticing things, thinking about things, wanting things, considering options, taking decisions, learning things, wondering whether, wondering why, trying to recall, becoming afraid then hopeful, and finally relieved.

But not only are there conscious processes of those types. Less obvious are the myriad *unconscious* information manipulating processes which underpin all that activity, many of them tightly integrated with processes which do control bodily processes. The unconscious processes fall into many different categories, some of them clearly mental processes, closely related to those of which we are conscious, for instance the unconscious processes involved in recognising words and grammatical structures when reading or listening to speech. Other processes are clearly physical rather than mental, for instance the firing of neurons, the manufacture of neurotransmitters, the transmission of chemicals from one part of the brain to another.

Different scientific disciplines and sub-disciplines, including linguistics, psychology, sociology, anthropology, psychophysics, psychiatry, neuroscience, biochemistry and physics seem to get a grip on different types of processes.

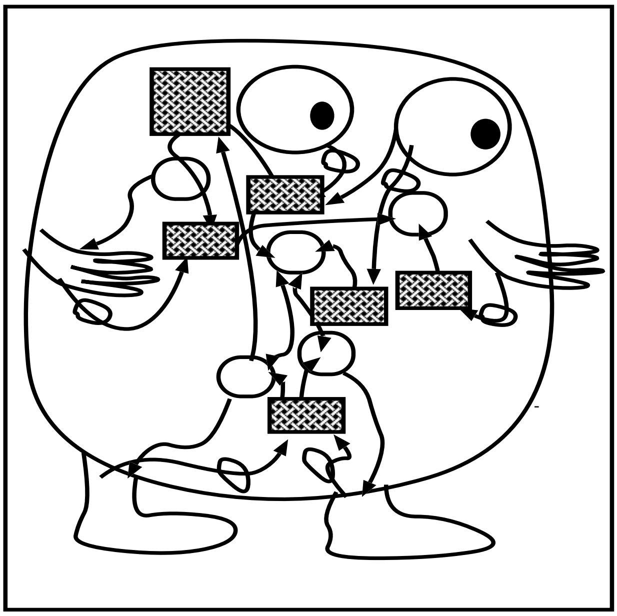

Figure 1: **Multi-process agents**
Here rectangles represent short or long term databases and ovals represent processing units. Arrows represent flow of information.

Is there any view which unifies them all?

In principle it could turn out that there is simply a vast collection of processes at many levels which just happen to produce the phenomena we are interested in explaining, but which are too messy and too complex to be understood by us. Such a view might be expressed as in Figure 1 suggesting an essentially "flat" architecture composed of very large numbers of interconnected processing mechanisms with no discernible overall structure.

2. ARCHITECTURES

If there is any hope of understanding our minds at different levels in any detail it is likely to make use of the notion of "architecture". An architecture is a system of interacting modules performing different functions, such as transducing sensory data, interpreting data, storing data, making inferences, generating new goals, resolving conflicts, learning patterns, storing generalisations, controlling various internal processes, initiating or controlling external actions, and so on.

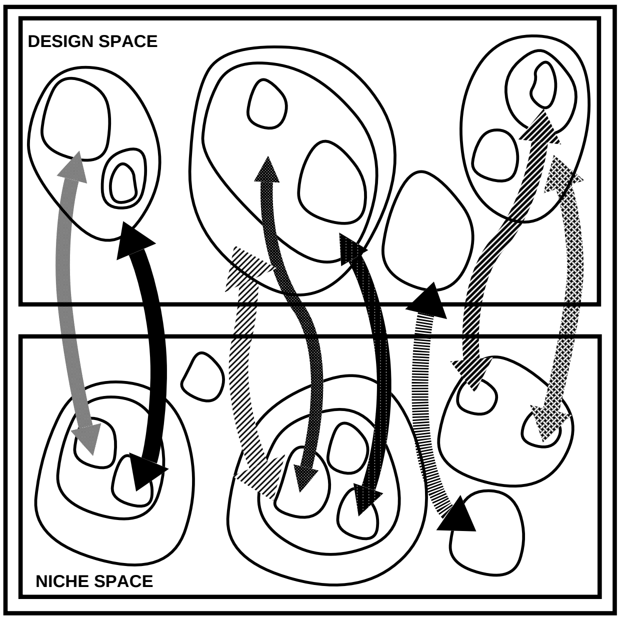

Figure 2: **Design space and niche space**
Arrows linking designs and niches depict different sorts of complex "fitness" relationships (usually involving tradeoffs). Changes in one design can alter the niche of another, which in turn can lead to design changes, which alter the niche of the first. Interacting trajectories in both spaces may involve multiple feedback loops.

The modules themselves may have architectures. Some modules may share components with others.

The architectures and their subdivisions need not be physical. Interacting architectures in different ontologies (different levels of abstraction) may coexist. For instance anatomists investigate the physical architecture of the brain and electronic engineers design the physical architecture of computers, whereas (some) psychologists study the functional architecture of the mind and software engineers design functional architectures for abstract (or virtual) machines like operating systems, compilers, spreadsheets, or networked file systems. These are not physical mechanisms even though they are *implemented* in physical mechanisms.

Virtual machines in brains and computers may have subtle and complex relationships with underlying physical or physiological mechanisms.

Can we understand human minds?
It's one thing to understand a complex system designed by people. The complexity produced by millions of years of evolution is a different matter. But if we have some idea how evolution achieves complexity then we can use the fact that minds are products of evolution to help us understand. This can augment other approaches, e.g. analysing functional requirements for the whole architecture or for components of the architecture; observing many kinds of performance in natural and artificial settings; charting the effects of various kinds of brain damage; comparing the capabilities and architectures of different sorts of animals; making use of ever more powerful non-invasive techniques for studying some of the physiological processes in living and active brains; and trying to design working systems with similar capabilities and learning from our failures.

3. TRAJECTORIES IN DESIGN SPACE AND NICHE SPACE

One way to understand evolution is think in terms of trajectories in design space and niche space. See Figure 2.

Varieties of adaptive, self-organising systems correspond to different regions of design space. Coexisting instances of various designs (and parts of designs) help to create niches for one another. The niches influence the trajectories of individuals as they learn, adapt and develop, thereby moving themselves through design space – sometimes crossing discontinuities. The niches also influence trajectories across generations, i.e. evolutionary developments.

When lots of interacting systems cooperate and compete while moving through both spaces we get multiple interacting trajectories in both spaces, generating new higher level feedback loops.

Perhaps this can help us understand biological evolution as a more self-directed process than suggested by the standard picture of a mixture of blind variation and selection only by success. Where some of the designs evolved are designs for virtual machines, fossil records are not going to help much, so a theoretical framework is essential if we are to fill gaps in empirical data.

4. TOWARDS AN ARCHITECTURE

Trying to achieve an understanding of how human minds and other minds work, requires a collection of parallel journeys with multiple themes leading through diverse worlds each with its own structures. Relevant ideas come from philosophy, psychology, AI, computer science, software engineering, social science, brain science and evolutionary biology.

One theme is the need to understand how abstract processes, like mental processes or virtual machine processes in software systems, relate to the underlying physical processes – part of the clue to resolving the alleged mystery of consciousness. The answer includes the important notion of circular causation between levels of reality: where low levels despite being causally complete in their own terms both cause and are are caused by processes at higher levels which are supervenient on them. Non-physical virtual machines are also involved in social mechanisms, e.g. when ignorance causes poverty and poverty causes crime.

Another theme (compare [4]) involves combining philosophical analysis of the presuppositions of many of our concepts used for describing ourselves (angry, humiliated, apprehensive, relieved, self-controlled, careless, attending to X's shape, attending to X's colour, attending to X's graspability, and many more), with an AI-oriented software engineer's attempt to work out the requirements

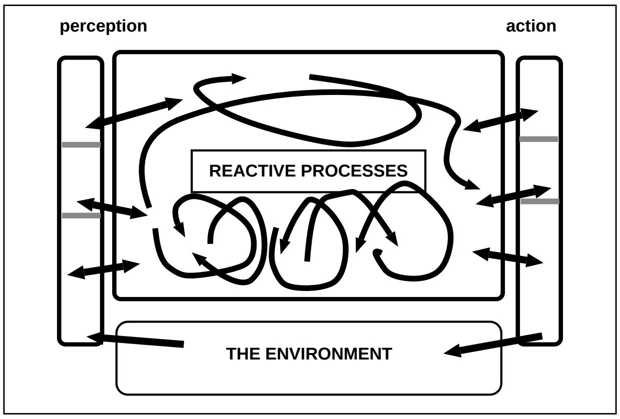

Figure 3: **A reactive architecture**
Complex examples include separate sensory and motor control subsystems, and may have many parallel hardware and "software" components all reacting to various combinations of internal and external events some of them modulated by internal state. Layered sensory and motor systems may deal with different levels of abstraction.

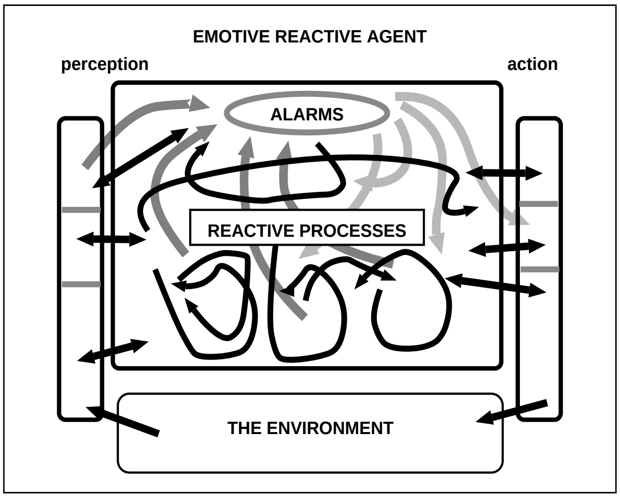

Figure 4: **Adding global alarms.**
Special fast-acting circuits can detect general patterns and trigger powerful reactions which dominate everything else.

for human-like robot (e.g. a versatile domestic servant and friend), to generate some "top-down" and "middle-out" design ideas for the architecture of a human-like mind.

Another requires learning from the empirical observations and theories of psychiatrists, psychologists and neuroscientists. Their work provides information about "bottom-up" design constraints, i.e. the sorts of processes that could possibly run on biological machinery, and also many clues about the human mental architecture based on empirical studies of how it works, and especially the interestingly different ways in which it can fail to work normally, after different kinds of brain damage or brain disease.

Such explorations over nearly three decades led me to a set of conjectures about the typical adult human virtual machine architecture. One of the main ideas is that different levels of sophistication developed at different times, and later levels did not replace earlier levels but coexisted with them, though usually in modified forms.

5. REACTIVE ARCHITECTURES

For many millions of years the only kinds of control architectures that existed, whether in single celled organisms or more sophisticated ones, were purely "reactive". Figure 3 crudely indicates a fairly sophisticated reactive architecture (perhaps an insect architecture) with a variety of sensors and motors along with a rich internal state which can be changed by various reactive processes, and which, in turn can trigger new changes or modulate changes triggered by other events, e.g. sensory input. The most important defining feature of a reactive system is that it cannot consider alternative hypothetical future sequences of actions, evaluate them and choose one. Thus it cannot create new plans, though if it has previously been designed or has evolved so as to include a set of stored plans it can be triggered to select one. New sequences of actions can be learnt by reinforcement of associations between conditions and subsequent actions, including internal actions. But this requires the actions to be *performed* so that positive or negative feedback can drive learning.

Conflicting reactions can be dealt with either by using some sort of vector addition or by making use of selection mechanisms (e.g. winner-takes-all neural nets).

Processes in reactive systems can operate at different time scales. Where some are relatively slow it may be necessary to have a global "alarm" system to produce rapid redirection, as in Figure 4. As so often in evolution this might be done by copying and modifying one of the pre-existing reactive modules. The modifications involved giving the module inputs from all over the system, making it work fast, and making it crudely classify inputs into categories relevant to certain global behaviours, e.g. freezing, fighting, fleeing, mating, becoming quiet and highly attentive, etc. Early versions could have purely innate categories and associations. Others might learn new ones using mechanisms like reinforcement learning.

Additional mechanisms could support primitive types of motivation. Some detected patterns (e.g. shortage of energy), instead of directly triggering behaviour might trigger a persistent state change which then helped to control subsequent behaviour (e.g. seeking food) until a need had been satisfied.

6. DELIBERATIVE MECHANISMS

The amazing diversity and success of insects shows how powerful reactive architectures can be. Such things as beehives and elaborate "cathedrals" built by termites show how reactive systems can produce cooperative behaviour with considerable functional differentiation in a whole population.

Reactive systems can achieve anything, provided that there has been time and opportunity for all the required sub-behaviours (plans) to be discovered in advance and "stored" either by evolution or in the development and training of individuals, and provided that there is sufficient storage capacity for all the potentially necessary associations between conditions and actions.

A human designer would consider avoiding the need for so much prior "experience" and so much storage, by adding a deliberative layer to the architecture, which can create evaluate and select sequences of actions in advance of performing them. This requires extensions to the reactive architecture, which might be achieved by a variety of modifications of the components already in a sophisticated reactive system. For instance the associative system which can learn which (internal) behaviour to produce in response to a particular state could be copied then modified to learn to predict what will be sensed in various contexts, or which behaviour would be produced by something else. I.e. it can become a *predictive* device.

Another copy of an associative reactive mechanism might be modified so that instead of producing an action it instead produces some sort of symbolic representation of that action, perhaps a copy of the types of signals to the motor subsystem that previously would have generated the action.

Combinations of such mechanisms might allow a system sometimes to generate a symbolic action sequence without performing the corresponding actions. This would allow the result of such a sequence to be evaluated, and a decision taken whether to re-do the actions "for real". As everyone knows, such anticipatory planning can allow disasters to be avoided.

This capability requires an important additional architectural feature, namely a re-usable memory in which the sequences can be constructed so that their consequences can be evaluated. Further developments could allow the memory to be used to construct more than one action sequence so that different options can be compared and one selected. These, along with other mechanisms, would combine to provide a deliberative extension to the original reactive type of architecture.

Moreover, just as purely reactive architectures sometimes need a fast global alarm system to take control where rapid action is urgently required, so might deliberative mechanisms: e.g. planning ahead could reveal a danger or opportunity requiring an urgent change of strategy. This could either use an extension of the original alarm system, or else a copy modified to react to the contents of the short term memory store, or the predictions of the associative memory, or some other combinations. A hybrid deliberative and reactive system with global alarm mechanism is sketched in Figure 5.

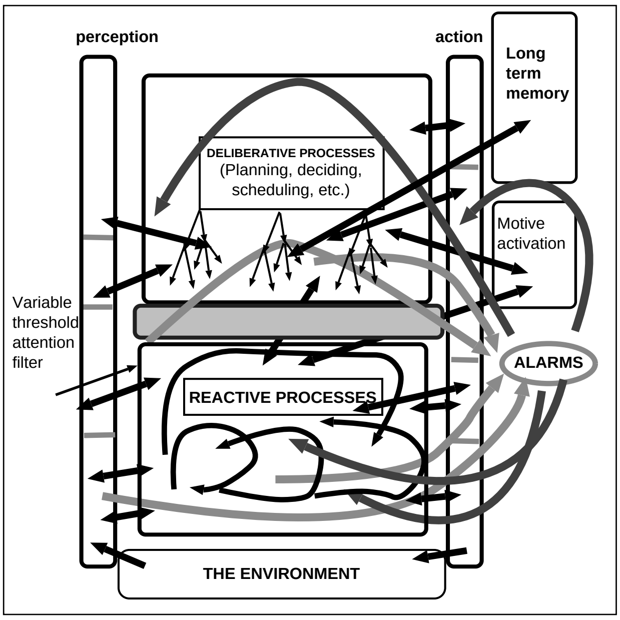

Figure 5: **A hybrid architecture with global alarms.** Reactive and deliberative mechanisms may sometimes be dominated by control signals from a global alarm system.

7. TWO SORTS OF EMOTIONS

The architecture so far sketched is beginning to be rich enough to be mapped (crudely) onto models developed in brain research. For example, Damasio's recent very influential book [2] makes a distinction between primary emotions which are triggered by external or internal stimulation of various sense organs, and secondary emotions which are triggered by purely cognitive events.

If we regard the global control signals produced by the alarm system (or alarm systems) as emotions, we see that in the context of a hybrid two layer system with alarm mechanisms there is support for both types of emotions. Our global alarm mechanism corresponds closely to the assumed role of the limbic system including the amygdala which is thought to learn associations of the type involved in emotions.(See also [6].)

In Damasio's theory secondary emotions *always* trigger the mechanism involved in the older primary emotion system, which produces a variety of physiological changes which, in turn, will be detected by various sensors, providing a characteristic "feel" which depends in part on the body state.

We can easily envisage a slight modification of the architecture where instead of the single alarm system whose outputs, even in response to secondary emotions, always go through the same physiological mechanisms there is another copy of the alarm system with a slightly different function. It is merely concerned with important and urgent cognitive responses and its outputs are restricted to react entirely within the deliberative (cognitive) mechanisms.

Alternatively the single global alarm system might develop so that for some kinds of states it produces none of the normal primary emotional outputs and merely causes high level changes in the deliberative mechanism. In humans, this sort of change, with emotional reactions becoming less physical and more central seems to be part of the process of growing up and becoming more emotionally mature [6]. It doesn't work that way for everyone however!

A "filter" (with a dynamically varying interrupt threshold [17]) is shown in Figure 5 since for some purposes the deliberative layer will be resource limited [10]. When performing urgent and important tasks it may need to be "protected" from interruptions coming either via the main reactive mechanisms (e.g. new motives demanding consideration) or via the global alarm system.

8. TOWARDS SELF-AWARENESS

In previous publications Luc Beaudoin, Ian Wright, Brian Logan and I have proposed that for certain purposes it will be useful if the deliberative processes in which internal actions are performed can be monitored, evaluated and perhaps modified ([1, 17, 13, 15]). We therefore proposed that in addition to the reactive and deliberative layers a further layer of functionality is useful: meta-management, as indicated rather sketchily in Figure 6.

The meta-mangement mechanisms could, for instance, be used to monitor deliberative processes and detect that certain strategies are more effective than others. There are other kinds of roles for which this could be useful, including monitoring a variety of internal states including intermediate processing states in sensory systems as happens when we notice, for instance, that our vision is blurred, or that the percept of an object viewed obliquely has a different shape in the visual field from the same object viewed from a different angle. These abilities are crucial to being able to produce realistic paintings, for instance.

I have argued elsewhere that if robots have this kind of ability and also (by recursive use of meta-management) notice that they have it, some of them may begin to reflect on the differences between the internal states which they detect in themselves and the perceived properties of physical objects in the environments. They would thereby have invented the idea of *qualia* and might easily be seduced by a host of standard philosophical arguments about the nature consciousness and its relationship to the underlying physical processes.

Another consequence of having a meta-management layer in the architecture is that the control decisions based on meta-management processes may sometimes be overridden by processes generated by the alarm system or other reactive mechanisms. For instance a person who is infatuated with someone, very jealous, very proud of his child's achievements, etc. may find it difficult to concentrate on tasks requiring close attention. Even deciding to concentrate may lead only to partial success: the pleasant or unpleasant thought, desire, memory or whatever gets through the "filter" and diverts attention. (This is elaborated in our discussion of grief in [17].)

Partial loss of control of attention and thought processes is a characteristic feature of human emotions that figure in plays, novels and social relationships. We could call them "tertiary emotions". They may, but need not, trigger processes in the primary emotion system, leading to changes such as sweating, tension, etc.

9. CONCLUSION

Within the sort of multi-layer architecture sketched here, we can begin to explain a host of familiar features of the human mind including the presence of both deliberative and reactive mechanisms, various kinds of learning, e.g. by positive and negative reinforcement, and by creation of new structures through the use of

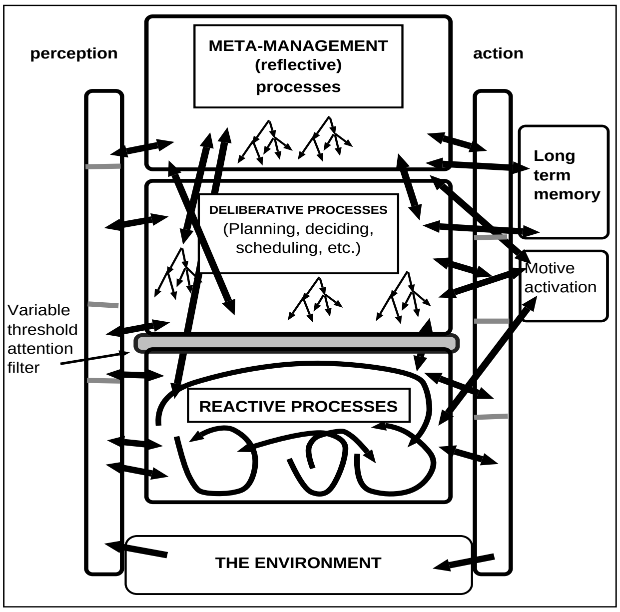

Figure 6: **A three layered agent Architecture**
(Note: global 'alarm' mechanisms not shown.)

deliberative mechanisms. The architecture also provides a framework for analysing various ways the system can go wrong. Many recent discoveries in brain science are compatible with these ideas. For instance mechanisms for global redirection are found in the brain stem and limbic system, especially the amygdala. Work by Damasio and others suggests that meta-management and aspects of deliberation work through frontal lobe mechanisms [2].

We also extend and unify various theories of emotions by accounting for three different types of emotional states related to control disturbances in the reactive, deliberative and meta-management layers (discussed further in some of the listed publications).

The architecture is not only plausible as regards the types of evolutionary processes that might have produced it, but it also begins to map onto new results from brain research. Moreover we can use it to criticise restrictive interpretations of such research e.g. by pointing out that in requiring secondary emotions *always* to trigger reactions involving the primary emotional mechanism Damasio is ignoring the important possibility of alarm mechanisms which sometimes operate *entirely* within the deliberative and meta-management layers, as seems to happen within emotionally mature individuals. (There is not space here for a full discussion.)

Examining an architecture with the kind of complexity discussed here shows that there are very many possible ways it can adapt, develop or learn, suggesting that current theories of learning and development are vastly oversimplified since they account for only a small subset of types of change.

We have also indicated how, within the sort of architecture proposed, some of the central phenomena which led to philosophical puzzles about consciousness can be explained. This explanation depends in part on a feature of the model which has not been described in detail here, though it is suggested in outline in the diagrams, namely that perceptual mechanisms are also layered, with different levels of analysis and interpretation proceeding in parallel and feeding data into different parts of the architecture. Then "qualia" will be related to the possibility of meta-management processes attending to some of the intermediate results of such sensory processing. This makes it possible to distinguish seeing the table in front of you and having an experience which may or may not be produced by an external object that looks like a table. (An incomplete paper on consciousness at my web site develops this point.)

Some researchers will object that this sort of model presupposes unnecessary "centralised" control. They espouse more distributed control, with many relatively independent processes taking local decisions out of which the appearance of coordinated intelligence emerges, as in the case of termites building their cathedrals.

It is not possible to rule out such models apriori. It requires an analysis of tradeoffs inherent in different architectures and the extent to which particular capabilities could or could not reliably emerge from a system with highly distributed control, along with the opportunities for such tradeoffs to influence trajectories in design space and niche space. My own view is that plants are the prime example of distributed control and the evolution of brains and nervous systems was a response to the need for more centralised control in mobile organisms, since otherwise different sub-mechanisms might choose to move in different directions, without a sensible resolution being possible (e.g. because simple voting schemes cannot cope with multiple sources of motivation.)

Is this sort of discussion relevant to the task of designing intelligent robots or software systems, or is it merely applicable to biological systems?

I believe that by considering the model in the context of trying to understand trajectories in design space and niche space we can see how although the features we have discussed are not merely the results of chance developments during the course of evolution, but also have some advantages over alternative designs. In particular, [14] and [9] both conjecture that mechanisms required for intelligence even in some artificial systems will also be capable of producing emotional states. However, not all the design features described here will be required in all such systems.

For instance, much of the human virtual machine architecture has to do with managing the complex physical machine in which it is implemented. For purely software agents, the requirements will clearly be different, though perhaps not as different as might be thought.

References

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

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

[3] Randall Davis. What are intelligence? and why? *AI Magazine*, 19(1):91–110, 1998. (Presidential Address to AAAI96).

[4] D.C. Dennett. *Kinds of minds: towards an understanding of consciousness.* Weidenfeld and Nicholson, London, 1996.

[5] Stan Franklin. *Artificial Minds.* Bradford Books, MIT Press, Cambridge, MA, 1995.

[6] Daniel Goleman. *Emotional Intelligence: Why It Can Matter More than IQ.* Bloomsbury Publishing, London, 1996.

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

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

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

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

[11] 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.

[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. What sort of control system is able to have a personality. In Robert Trappl and Paolo Petta, editors, *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Springer (Lecture notes in AI), Berlin, 1997. (Originally presented at Workshop on Designing personalities for synthetic actors, Vienna, June 1995).

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

[15] A. Sloman and B. S. Logan. Architectures and tools for human-like agents. In Frank Ritter and Richard M. Young, editors, *Proceedings of the 2nd European Conference on Cognitive Modelling*, pages 58–65, Nottkingham, UK, 1998. Nottingham University Press.

[16] Norbert Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine.* The MIT Press, Cambridge, Mass.:, 1961, 2nd ed.

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