**In: Proceedings 5th Scandinavian Conf. on AI, Trondheim May 1995,
IOS Press, Amsterdam, 1995**

Exploring design space and niche space

Aaron Sloman
School of Computer Science, The University of Birmingham, UK
A.Sloman@cs.bham.uk.ac

Abstract. Many who give definitions of AI offer narrow views based either on their own work area or the pronouncement of an AI guru about the scope of AI. Looking at the range of research in AI conferences, books, journals and laboratories suggests something very broad and deep, going beyond engineering objectives and the study or replication of human capabilities. This exploration of the space of possible designs for behaving systems (design space) and the relationships between designs and various collections of requirements and constraints (niche space) is inherently multi-disciplinary, and includes not only study of architectures, mechanisms, formalisms, inference systems, and the like (aspects of natural and artificial designs), but also the attempt to characterise diverse behavioural capabilities and the environments in which they are required, or possible. 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.

1. Introduction

Many AI researchers and writers of text books (one exception being [14]) think of AI as primarily a branch of engineering: an attempt to make machines that can perform difficult tasks, some of which could previously be performed only by human beings, and possibly some which not even human beings can perform. A striking example of the engineering viewpoint is McCarthy's recent paper [11] on giving robots self-consciousness, where he is at pains to distinguish his objectives from those of modelling or replicating the human mind. His is an important and worthwhile activity, and although it can and probably will contribute to the broader activity, that is not his main objective.

There is a different view of AI as a very broad field of study, to which the engineering activities can contribute as a subfield, and in which engineering techniques are crucial, though the objectives are not all practical. This more general view of AI is not merely my subjective preference, but can be abstracted from the range of topics and research activities to be found in AI conferences, AI journals and AI laboratories. It can be roughly characterised as:

> 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 at some time in the future).

2. The design-based approach to the study of mind

From this viewpoint, AI is the exploration of the space of possible designs for partial or complete agents, and overlaps with Alife.[1] This is more general than cognitive science, which

[1] This is one of several papers from the Birmingham 'Cognition and Affect' group, presenting the design-stance view of a mind as a sophisticated self-monitoring, self-modifying control system: [25, 13, 5, 16, 17, 18, 19, 20, 28, 21, 22, 23, 24, 29]

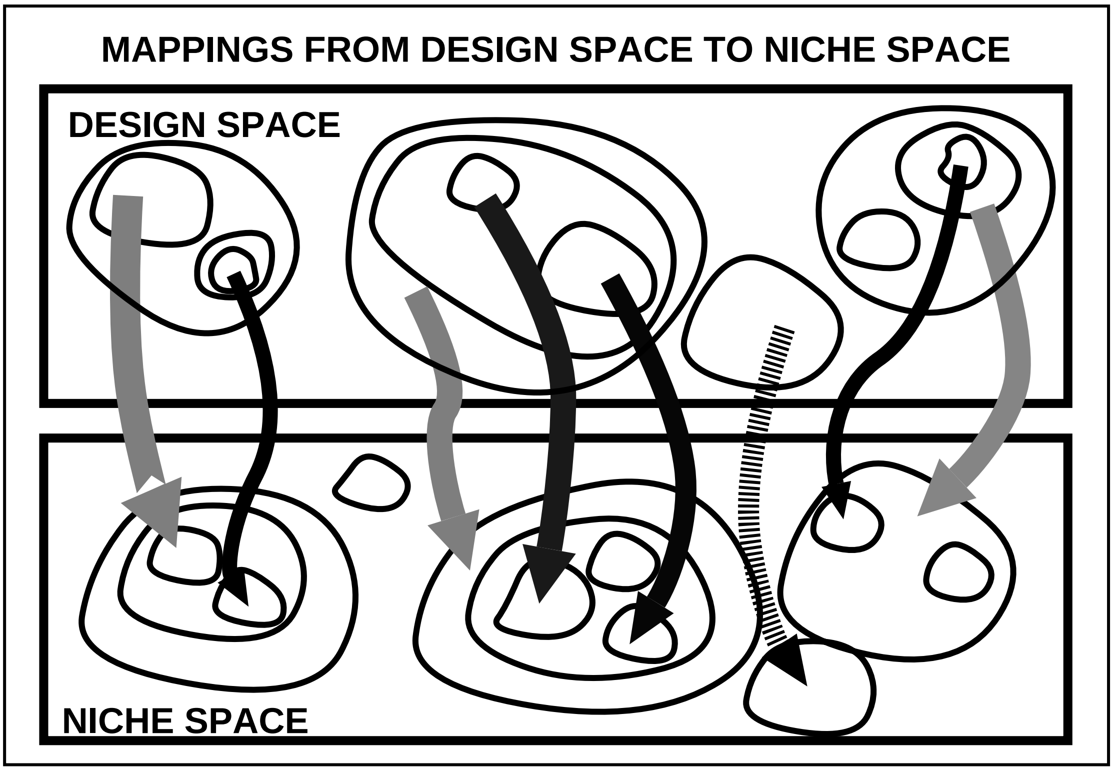

Figure 1: **Mappings between design space and niche space**

(A niche is a set of requirements. Mappings can vary in degree and kind of goodness.)

is normally restricted to the study of humans and other animals (e.g. [9]). Gardner's book ([7]) postulates multiple forms of intelligence but restricts itself to forms that already exist. Minsky's approach in [12] is closer to ours.

The work of AI engineers is obviously relevant to this broad study insofar as they discover techniques, concepts and principles that extend ideas that originate in laboratory or other investigations of existing intelligent agents, whether humans or other animals. Engineering ideas often play a useful role in understanding natural systems. An old example of a concept relevant to both engineering and biology is 'feedback' (see [27]), though there are many other cases. For example the task of designing aeroplanes has led to a deep understanding of principles of aerodynamics and design tradeoffs that are also relevant to understanding the differences in designs and capabilities of different birds.

Conversely, the study of natural intelligence, in humans and other animals, can give us ideas that are relevant to the design of useful artefacts. This is particularly important when the artefacts are intended to communicate with humans, act on human goals, use human criteria for resolving conflicts and to deal with the unexpected in ways that are acceptable to humans.

3. Design space and niche space

One reason why the space of *possible* designs is so important is that no one design, whether natural or artificial, can be understood fully if we don't know what difference it would have made had the design been different in various ways. Full understanding, therefore, requires analysis of similarities and differences between actual and possible designs. However, there is more to be studied than designs: designs fit into environments and tasks. The biologists' notion of a 'niche' can be generalised as the notion of a set of requirements. Understanding a design is in part a matter of understanding how it relates to a niche and understanding how

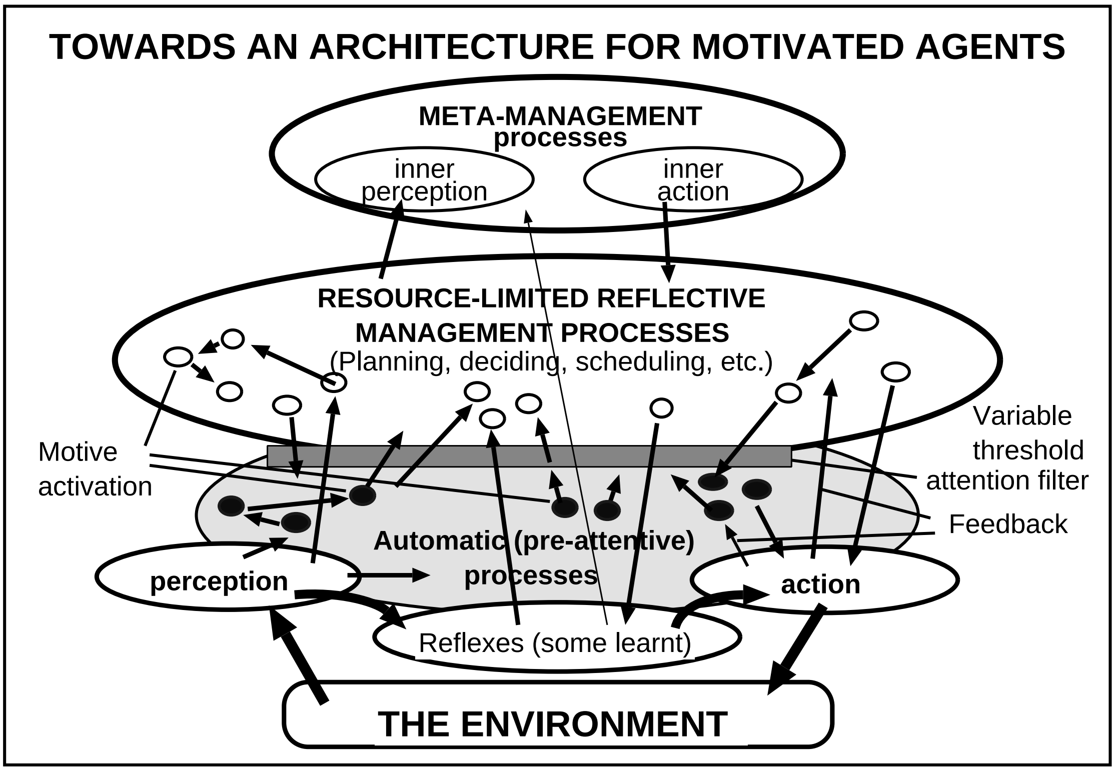
TOWARDS AN ARCHITECTURE FOR MOTIVATED AGENTS
META-MANAGEMENT processes
inner perception
inner action
RESOURCE-LIMITED REFLECTIVE MANAGEMENT PROCESSES
(Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
Feedback
Automatic (pre-attentive) processes
perception
action
Reflexes (some learnt)
THE ENVIRONMENT

Figure 2: **Towards an Intelligent Agent Architecture**

changing the design would change the niches that it fits well. An agent's niche is not just its physical environment: it includes, for instance, capabilities of other agents, of the same and different types. Niches can evolve too. As designs and niches become more complicated they correspond to smaller regions in the diagram, though a new design feature may be relevant to a large region in niche space.

We need to explore the mappings between design space and niche space indicated approximately in Figure 1. The figure oversimplifies in many ways, not least because there is not just one design space nor a single niche space: both are describable at many levels of abstraction. My colleague, Riccardo Poli, has suggested that one can see both diagrams as horizontal slices through a terrain of varying contours. At a lower level there may be more primitive and general mechanisms, applicable to a wide range of niches. At a higher level both designs and niches become more specialised. In humans there are clearly many coexisting levels of design: we share many capabilities with organisms that evolved much earlier, but we can control, modulate, and deploy those abilities in far more ways (cf. [6]).

4. Architectures not algorithms

It is argued in [18] that the search for powerful explanatory architectures is more important than the search for algorithms. Figure 2 indicates very loosely a sketchy high level design for an architecture for a human-like intelligent agent, with many cooperating subsystems, some of which, in the grey area, operate in a very 'automatic' fashion, whereas others, labelled as 'management processes' involve explicit consideration of alternatives, creation and evaluation of options, and selection. A major difference between the two is that the management processes need a representational capability which includes the power to represent things that do not exist, e.g. possible futures and the actions that are not selected. This in turn is part of a more general requirement to be able to store information for future use, to postpone action, to abandon something strongly desired in favour of other more urgent or more important goals,

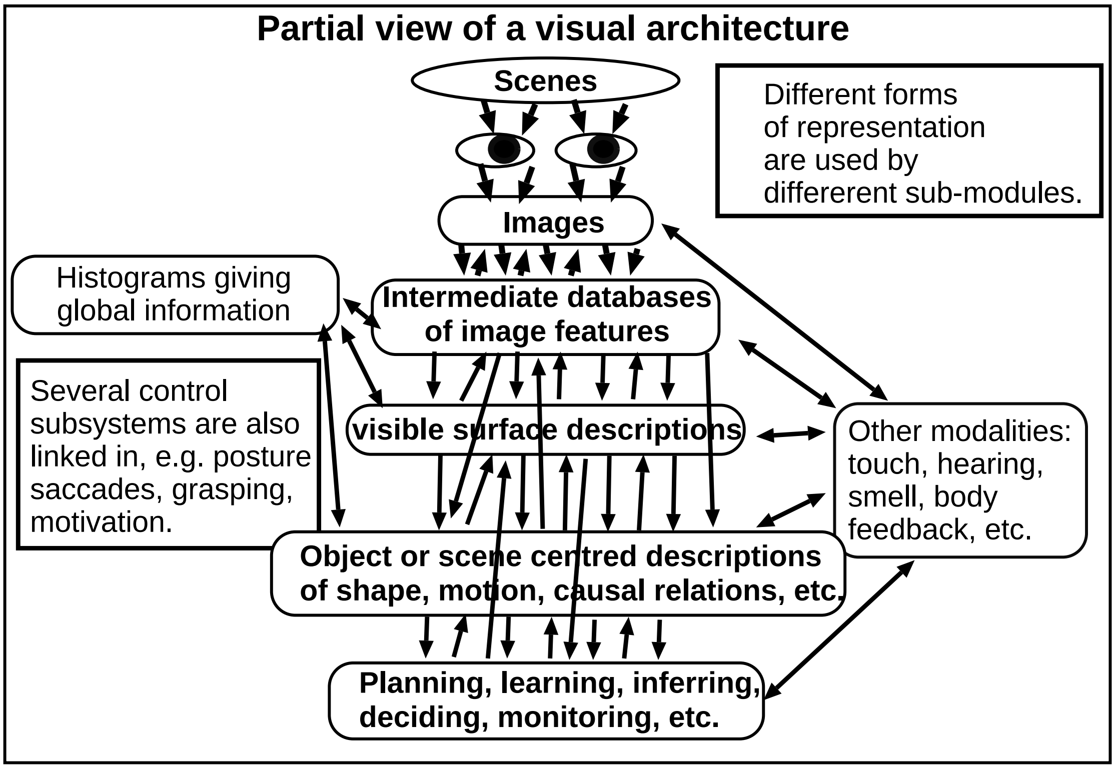

Figure 3: Sketch of a visual sub-architecture

and so on. We are exploring such an architecture as a possible design solution to a collection of requirements derived in part from analysing aspects of human mental life [5, 17].

As an illustration of the ways in which Figure 2 is too abstract, consider the detail missing with regard to perception. An analysis of requirements for a perceptual architecture can be found in [16], based in part on the ideas of Gibson [8], and suggesting the need for something of the sort crudely indicated in Figure 3, which is intended to indicate many different levels of analysis of interpretation, different processes simultaneously operating with different types of representations, processing information concerned with different subject matter, e.g. low level image features, more global image features, static features, dynamic features, visible surfaces in the scene depicted by the image, other geometrical aspects of the scene (e.g. hidden surfaces), causal relationships, other relationships, and more besides. This partly overlaps with Marr's theories [10], though he seemed to think the system would be more modular and less labyrinthine. There are many other gaps in the diagrams.

In order to make progress, it is often useful temporarily to ignore some of the details, as part of the process of understanding more global designs and design requirements. That involves a study of what Bates and colleagues ([3]) call 'Broad and shallow architectures'. From an engineering standpoint these may be ends in themselves. From our standpoint they are merely a useful transitional stage, while we struggle to find more of the missing parts of the jigsaw puzzle.

Another view of the architecture is hinted at in Figure 4, which is intended to suggest dynamic aspects of the design, (compare [15], [17]), namely the existence of different sorts of control states, at different levels of control, with varying types of influence, different life-spans and different degrees of ease of change. An exploration of design space would include comparing (a) control hierarchies based only on the sorts of feedback loops which control engineers study using partial differential equations involving a fixed number of continuous variables, with (b) control architectures that include changing structures with

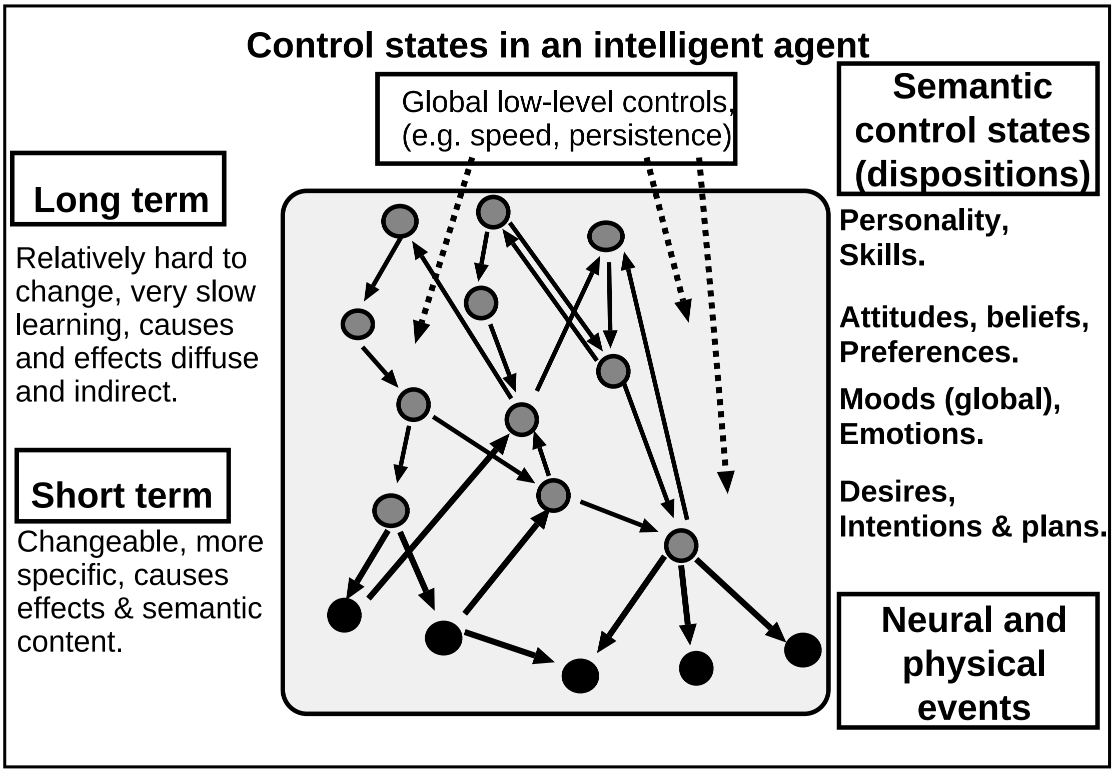
Control states in an intelligent agent
Global low-level controls, (e.g. speed, persistence)
Semantic control states (dispositions)
Long term
Relatively hard to change, very slow learning, causes and effects diffuse and indirect.
Short term
Changeable, more specific, causes effects & semantic content.
Personality, Skills.
Attitudes, beliefs, Preferences.
Moods (global), Emotions.
Desires, Intentions & plans.
Neural and physical events

Figure 4: Control states of various kinds

varying complexity, such as plans and parse trees, some of whose changes are discontinuous. For control mechanisms of type (b), new kinds of mathematics may be needed.

Our own work includes trying to see how much of that control hierarchy can be based on a broad and shallow design of the sort indicated in Figure 2, in which the higher level processes are far more like symbolic AI processes and the lower level, pre-attentive, processes are more like spreading activation processes in networks. Both sorts could include a mixture of symbolic and sub-symbolic processes.

5. Confusions to be avoided

There are many complications and many ways of misinterpreting this programme. For example talk of studying 'design' can mislead people into thinking that the methodology involves working only top down, from requirements through specifications to implementations. That is not so, for designs can be studied and related to various niches, no matter how they were discovered. Some are discovered by working up from lower level mechanisms, some by exploring natural systems and some by trying to create new engineering solutions.

Talk of 'design' may suggest a concern restricted to practical goals and useful artefacts. That is not so. A design is an specification for a working system, and understanding designs, or classes of designs, is as much a part of general science as understanding physical laws and their implications. People often fail to appreciate that there is a concept of 'design' which is equally applicable to natural and to artificial phenomena: a design, in the sense intended here, may be found in a plant or animal, even though there was no designer, at least no designer with explicit goals, for evolution can be thought of as a designer.

Another common mistake is to assume that the design-based approach aims to find a single design. The emphasis on *spaces* of designs and niches above, should explain why that is too narrow. It must be admitted, however, that much work in AI merely aims to produce a single design, and does not explore alternatives and analyse their impact on mappings into niche space. Sometimes this reflects practical concerns or funding limits, and sometimes narrowness of vision.

Yet another confusion is based on the assumption that architectures must be physical. That is not so, for much of the work of software engineers is concerned with designing 'virtual' or 'abstract' machines (e.g. software architectures) that are not physical, but are implemented in physical systems. The relationships between virtual and physical components need not be simple. For instance there may be far more 'parts' in a large sparse array in a virtual machine than components in the physical implementation. This is a complex notion, and a full discussion of it would lead into the analysis of the philosopher's notion of supervenience, and an analysis of the concept of causation: for I claim that events and processes in virtual machines can have causal powers. (This is discussed further in [21]. See also [2] and [26].) This paper cannot give a full account of the methodology: the bibliography includes a selection of books and papers that implicitly or explicitly expand the ideas.

6. Conclusion

The design-based approach contrasts with the 'blindly empirical' approach of many psychologists, who merely seek correlations between observables, and with Dennett's ([1]) 'intentional stance', which, like Newell's 'knowledge level analysis', has to presuppose that agents are rational. One consequence of the design-based approach to the study of mind is that within the framework of an architecture we can generate precisely defined concepts describing many kinds of states and processes that can occur at the information processing level, just as development of a theory of the architecture of matter generated new concepts of types of *stuff*. In both cases, the new concepts will overlap to some extent with but will be far more extensive, precise, and systematically extendable than ordinary pre-theoretic concepts. Thus familiar notions like: 'think', 'imagine', 'conscious', 'desire', 'enjoy', 'emotion', 'pleasure', 'pain', 'personality', and many more, will either be shown to have a basis in the architecture, or will be replaced by new architecturally grounded concepts. We shall then also be in a much better position to discuss which of the concepts are and which are not applicable to other organisms and machines with different architectures, and to ask more sharply defined questions about evolution. The implications of all this are profound: e.g. for engineering, for biology, for psychology, for therapy, for education, for philosophy, and for our view of how we fit into the scheme of things.

References

[1] Dennett, D.C. (1978) *Brainstorms* Bradford Books and Harvester Press

[2] Dennett D.C. (1991). *Consciousness Explained* Allen Lane, the Penguin Press.

[3] Bates, J., Loyall, A. B., & Reilly, W. S. (1991). Broad agents. Paper presented at the AAAI spring symposium on integrated intelligent architectures. Stanford, CA: (Available in SIGART BULLETIN, 2(4), Aug. 1991, pp 38-40.).

[4] Beaudoin, L.P & Sloman, A (1993) A study of motive processing and attention, in A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay (eds) *Prospects for Artificial Intelligence* IOS Press, pp 229-238

[5] Beaudoin, L.P (1994) *A design-based study of autonomous agents* PhD thesis, School of Computer Science The University of Birmingham.

[6] Brooks, R.A. (1991) Intelligence without representation *Artificial Intelligence* 47, 139-159.

[7] Gardner, Howard (1985) *Frames of Mind: The Theory of Multiple Intelligences* Paladin Books, London. (Originally Heinemann 1984).

[8] Gibson, J.J. (1979) *The Ecological Approach to Visual Perception* Lawrence Earlbaum Associates, 1986 (originally published in 1979).

[9] Johnson-Laird, P. N. (1989) *The computer and the mind: An introduction to cognitive science,* Fontana (Second edition 1993)

[10] Marr, D. (1982) *Vision*, Freeman

[11] McCarthy, John, (1995) Making Robots Conscious of their Mental States, in *AAAI Spring Symposium on Representing Mental States and Mechanisms* 1995, Stanford, accessible via http://www-formal.stanford.edu/jmc/

[12] Minsky, M.L. (1987) *The Society of Mind* London: William Heinemann Ltd.

[13] Read, T. & Sloman A, (1993) 'The terminological pitfalls of studying emotion' in Read, T. (ed) *Proceedings WAUME93: Workshop on Architectures Underlying Motivation and Emotion, Aug 11-12 1993* Cognitive Science Research Papers, School of Computer Science, The University of Birmingham.

[14] Russell, Stuart & Norvig, Peter (1995) *Artificial Intelligence, A Modern Approach* Prentice Hall

[15] Simon, H.A. (1979) 'Motivational and Emotional Controls of Cognition' 1967, reprinted in *Models of Thought,* Yale University Press, 29–38

[16] Sloman, A (1989) 'On designing a visual system: Towards a Gibsonian computational model of vision', in *Journal of Experimental and Theoretical AI* **1,4**, 289-337

[17] Sloman, A (1992a), 'Prolegomena to a theory of communication and affect' in Ortony, A., Slack, J., Stock, O. (Eds.) *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, Heidelberg, Germany: Springer, 229-260. (Also available as Cognitive Science Research Paper, School of Computer Science, University of Birmingham.)

[18] Sloman, A. (1992b) The emperor's real mind: review of Roger Penrose's *The Emperor's new Mind: Concerning Computers Minds and the Laws of Physics*, in *Artificial Intelligence* 56 pp 355-396, (Also Cognitive Science Research Paper, University of Birmingham)

[19] Sloman, A. (1993a), Prospects for AI as the General Science of Intelligence, in A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay (eds) *Prospects for Artificial Intelligence,* IOS Press, pp 1-10.

[20] Sloman, A. (1993b), The mind as a control system, in *Philosophy and the Cognitive Sciences*, (eds) C. Hookway & D. Peterson, Cambridge University Press, pp 69-110

[21] Sloman, A (1994a) Semantics in an intelligent control system, in *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering* Vol 349, 1689, pp 43-58

[22] Sloman, A (1994b) Computational modeling of motive-management processes, *Proceedings of the Conference of the International Society for Research in Emotions*, Cambridge, July 1994. Ed N.Frijda, p 344-348. ISRE Publications.

[23] Sloman, A. (1994c), Explorations in Design Space. *Proceedings 11th European Conference on AI* Amsterdam 1994.

[24] Sloman, A. (1995 to appear) Towards a general theory of representations, in D.M.Peterson (ed) *Forms of representation* Intellect press

[25] Sloman, A & Croucher M. (1981) Why robots will have emotions, *Proc 7th Int. Joint Conf. on AI,* Vancouver

[26] Taylor, C.N, 1992 *A Formal Logical Analysis of Causal Relations* DPhil Thesis, Sussex University. Available as Cognitive Science Research Paper No.257

[27] Wiener, N. 1948, revised ed. 1961, *Cybernetics: or Control and Communication in the Animal and the Machine* 2nd ed. Cambridge, Mass.: The MIT Press (1st edition 1948)

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

[29] Wright, I.P, Sloman, A, & Beaudoin L.P (To appear.) The architectural basis for grief, presented at *Geneva Emotions Week* 8-13 April 1995. Draft version available as ftp://ftp.cs.bham.ac.uk/pub/dist/cog_affect/geneva.ps.Z