In *Prospects for Artificial Intelligence*, Proc. AISB'93, (1993)
Eds. A. Sloman, D. Hogg, G. Humphreys, D. Partridge & A. Ramsay,
pp. 229–238, IOS Press, Amsterdam

# Introduction by Programme Chair: Prospects for AI as the General Science of Intelligence

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

**Abstract.** Three approaches to the study of mind are distinguished: semantics-based, phenomena-based and design-based. Requirements for the design-based approach are outlined. It is argued that AI as the design-based approach to the study of mind has a long future, and pronouncements regarding its failure are premature, to say the least.

## 1. Introduction

The death of Artificial Intelligence, or more specifically the failure of the computational approach to the study of mind, is often announced, either by converts to some 'new' approach, or by people who don't like the idea of a scientific or mechanistic explanation of how the human mind works. Such announcements are naive insofar as they assume that we already know (a) what can and cannot be explained or modelled in computational terms, and (b) what we are trying to explain or replicate. The first assumption ignores the fact that the study of computation is still in its infancy. The second assumption ignores the depth of our ignorance about the nature of the human mind.

Regarding (a): During its short life computer science has deepened our understanding of which sorts of processes are possible and useful, and which are not. For example it has produced a sequence of generalisations of the notions of mechanism and process, through the development of new forms of 'programming' language; and, in studies of 'complexity' issues, it has explored the boundary between what is theoretically or mathematically possible and what is feasible in the physical universe. Yet we still have a very long way to go, for instance in exploring forms of parallel computation, and the huge variety of virtual machines that have properties quite unlike the physical machines in terms of which they are ultimately implemented.

Regarding (b): although most people think they understand the essence of mind from direct experience of their own, much of this confidence is misplaced, and the concepts of ordinary language used to formulate questions or theories are nearly all ill-defined and unsuitable for the task. Thus people who think they know what questions they are posing when they ask what the function of consciousness is, how consciousness evolved, or whether consciousness can be explained in computational terms, are probably deceiving themselves, as we'll eventually discover.

By comparison with the nature of matter, our understanding of the nature of the human mind is pathetically shallow and incomplete, and worse, riddled with confusions. Where there's conceptual confusion, we can expect futile arguments at cross-purposes and unproductive investigations. Collection of new data from experiment and observation on humans and other animals is an endless process, but unless driven by deep theory-based concepts it does not in itself provide insights into the nature of the underlying mechanisms. If pursued in the light of good theories about mechanisms, new data can help to expose flaws in our hypotheses and drive the development of new theories that engage better with the fine structure of reality.

At least three different activities need to be inter-related: (i) exploration of new forms of mechanisms including new forms of computation, (ii) the construction of working models using those mechanisms, and (iii) collection of significant empirical data. All three need to go hand in hand, each feeding off, stimulating and correcting the others. This is what AI as the science of intelligence is all about. Alas, too often people pursue narrow interests tied to only one or two of these three sub-disciplines of the study of intelligence.

I include under the umbrella of AI several approaches that put themselves forward as rivals to symbolic AI, including connectionism, genetic algorithms and artificial life models. It is silly to treat these as mutually exclusive rivals in the current state of our knowledge. For all we know now, it could turn out that intricate combinations of these techniques, or their successors, will be required in the end, both for explaining human capabilities and for building useful human-like engines. As indicated below, it may also be sensible for AI to take account of some processes that many people would not class as computational, e.g. chemical processes such as the effects of alcohol and other drugs.

## 2. Work to be done

There are many aspects of human and animal competence that we cannot yet explain or simulate. In the early days of computers it seemed to some that the ability to do logic and mathematics or play 'intellectual' games like chess were going to be the main challenges for the study of intelligence, since these were things people found hardest. But we've since learnt that there are ways of getting machines to match or exceed normal human performance in sub-areas of mathematics, chess, and other similar activities, without necessarily imitating human ways of doing things in any depth. By contrast there are capabilities that are apparently achieved effortlessly even by children and other animals but which we cannot yet begin to approach on computers. E.g. although there is a considerable amount of AI work in vision inspired by both scientific and engineering goals, existing systems come nowhere near the generality, flexibility, and speed of biological vision (a topic developed further in Sloman [19]). (Though some *special-purpose* AI image processing systems may perform a specific task better than humans.) Moreover, phenomena associated with various forms of brain damage show that visual competence is composed of sub-competences that can exist or degrade independently in ways that do not correspond at all to the organisation of visual competence in current AI models of vision. Similar criticisms of the current state of AI compared with human abilities can be made concerning sub-fields such as language understanding and production, problem solving, planning, learning, and motor control. Even when we look closely at human performance in tasks where computers already do as well as or better than people we find striking differences in how those abilities are learnt and how they integrate with other abilities.

Although AI's achievements hitherto lag a very long way beyond its long term objectives, that in itself is not a reason for pessimism about its prospects. The work of people like Galileo and Newton lagged far behind the long term objectives of physics. The viability and vitality of a discipline is not defined by the successes and failures at any particular time, so much as by the variety of unsolved problems to which the approach of that discipline seems better suited than rivals at that time. Right now I do not know of any other discipline that provides tools and techniques better suited than those of AI to approach the explanation of intricate information-processing capabilities that characterise the human mind. Theories within other disciplines that study humans often lack the ability to account simultaneously for the fine structure and global organisation of human abilities, even though they may help to identify what needs to be explained.

Like the concepts and techniques used by Galileo, current AI tools and techniques may not yet have the power required for long term success. Nevertheless the disciplined, relentless, critical application of the most promising techniques available is a sure way to discover requirements for their successors: that's how science and engineering make progress.

## 3. Approaches to the study of mind

Different varieties of AI have in common a *design-based* approach to the study of mental phenomena, which contrasts with *semantics-based* and *phenomena-based* approaches to the study of mind. The three are not mutually incompatible: some researchers combine two or more, though this is rare.

### *3.1. Semantics-based approaches to the study of mind*

Semantics-based theories attempt to find out how people interpret words and phrases of ordinary language describing mental states and processes: they explore the structure of some portion of the lexicon of ordinary language (e.g. Ortony et al. [11]; Clore & Ortony [4]; and Johnson-Laird & Oatley [8], unlike Johnson-Laird [7], which presents a design-based theory). As a source of information about mental processes such enquiries restrict us to current 'common sense' with all its errors and limitations. Some of the conceptual analysis techniques of philosophers also produce semantics-based theories, though most eschew empirical surveys (e.g. Ryle [13], Sloman [16] chapter 4) and sometimes they are more concerned with the space of *possible* concepts rather than simply analysing *existing* concepts.

### *3.2. Phenomena-based approaches to the study of mind*

Phenomena-based investigations abound in the work of many psychologists. Such researchers presuppose that we already understand clearly what we are talking about when we refer to some phenomenon, such as "consciousness", "emotion", "motivation" and that examples can be intuitively recognized (e.g. emotional states). They then investigate other things that are correlated with the phenomenon in some way, e.g. environmental causes, physiological causes, physiological effects, behavioral responses, cognitive processes. The correlations may or may not include experimentally manipulated variables. (I shall not here discuss the vexed question as to whether correlations demonstrate causal connections.)

Phenomena-based theories that *appear* to be concerned with mechanisms, because they relate behaviour to neurophysiological structures or processes, often turn out on close examination to be concerned only with empirical correlations between behaviour and internal processes: they do not show *why or how* the mechanisms identified produce their alleged effects. That requires something analogous to a mathematical proof, or logical deduction, and most cognitive theories fall far short of that.

### *3.3. Design-based approaches to the study of mind*

The design-based approach attempts to go beyond these limitations by adopting the 'design-stance' (Dennett [1]). Designs are not restricted to artificial systems: we can analyse the design of a biological organism by (i) investigating its capabilities and the constraints within which it has to function, (ii) explaining how it is enabled to meet these 'requirements' by its architecture and the mechanisms used, and (iii) showing how these features imply satisfaction of the requirements. The last is akin to a logical or mathematical demonstration. The concept of "design" used here is very general, and does not imply the existence of a designer. Neither does it require that where there's no designer there must have been something like an evolutionary selection process. We are not primarily concerned with *origins*, but with what underlies and explains capabilities of a working system.

A full design-based explanation is rarely achieved in practice. It is quite a complicated affair, ideally including five different sorts of components: analysis of requirements, high level design specification, implementation details, analysis of how the design meets the requirements, and exploration of alternatives and trade-offs. These points are spelled out in more detail in (a) to (e) below, and later sections. There are many varieties of design-based approach, but what they have in common is that they all directly or indirectly contribute towards construction of a multi-faceted theory that includes all the elements I have mentioned. Many design-based theories do not include all five components, and are therefore partly lacking in explanatory power. The five components cannot be sharply distinguished since they overlap, and are to some extent recursive, as will become clear. The first three, (a), (b) and (c) below, are commonplace among engineers, and correspond approximately to Marr's three levels [9]. They are not enough for science: (d) and (e) are also needed. (e) is also useful when engineers need to evaluate design options.

*(a) Analysis of requirements for an autonomous intelligent agent.*
The requirements include: relevant features of the environment with which the agent interacts, resource constraints within the agent, whether the agent is part of a social system or not, which sorts of behaviours the agent should exhibit, and so on. The relevant features of the environment may be subtle and difficult to identify. For example, it is not obvious which aspects of the environment a human-like visual system needs to be able to detect and distinguish. If the agent interacts with other intelligent agents, the requirements will have to include the forms of interaction, and if that includes linguistic communication will need to specify the type of language, its syntax, semantics, pragmatics, etc. Often people who criticise AI by showing how human abilities differ from AI models are thereby unwittingly helping to refine and clarify the requirements to be satisfied by design-based theories.

*(b) A design specification for a working system meeting the requirements in (a).*
This should include an architectural analysis, i.e. a global decomposition of the system into major components with a description of their functional relationships and interfaces. A design can be recursive: it may include replication of points (a), (b), (c), (d) and (e) for some of the components, i.e. providing detailed functional requirements, a specification of how to meet those requirements in terms of a sub-architecture, details of how the architecture is to be implemented, and theoretical analyses of types (d) and (e) for the components. The design specification may be restricted to a single high level virtual machine, or it may include specifications at lower implementation levels, in which case it overlaps with (c).

*(c) A detailed implementation or implementation specification for a working system.*
Depending on the objectives of the investigation this may be intended merely as a simulation with predictive power, or as a realistic model, down to some level of detail. In either case, the implementation may involve several levels of virtual machines and possibly also a number of partly independent, asynchronously interacting, sub-systems. The virtual machines at different levels may be very different in character, for instance some constituting symbol-manipulating machines, some connectionist machines, and some electro-chemical engines.

In some architectures there can be close coupling between low level and high level virtual machines, making the models extremely hard to analyse: for example a machine that allows high level programs to alter its own microcode, or which contains microcode routines that invoke high level procedures. It may turn out that some chemical processes in the brain also illustrates such close coupling, for instance when alcohol, drugs or hormones alter high level qualitative behaviour whilst cognitive processes alter chemical processes.

*(d) Theoretical analysis of how the design specification and the implementational details ensure or fail to ensure satisfaction of the requirements.*
We must allow for requirements that cannot all be satisfied perfectly, so that some viable designs merely approximate to the requirements. The analysis, sometimes referred to as 'design verification', generally requires logical or mathematical formalisation of the relationships between (a), (b) and (c) and is often very difficult. The proofs may need horrendous combinatorial complexity in order to cover all cases, making them totally impractical in reality, in which case approximate intuitive analysis combined with systematic testing is all that can be achieved.

*(e) Analysis of the neighbourhood in 'design-space'.*
A full design-based theory would locate human mechanisms within a 'space of possible designs,' covering both actual and possible organisms and also possible non-biological intelligent systems (artifacts). Considering implications of possible alternatives to a particular design **D** leads to deeper understanding of that design. For example, demonstrating the consequences of replacing or modifying some feature **F** of **D**, helps to reveal the significance of **F** in **D**. More generally, studying contrasts between design options helps us understand trade-offs within any one design. This is related to (d) insofar as a rigorous mathematical analysis of a *particular* design or implementation can provide a basis for showing how variations in the design or implementation relate to variations in requirements, environments and constraints. (Compare Grossberg [5].) So work on (d) contributes to analysis of type (e). Moreover, (e) is essential if (d) is extended to showing that a particular design is *optimal*.

One benefit of either a *working* system implementing a design-based theory, or a mathematical analysis to show relationships between system features and behavioural capabilities, is that it can sometimes demonstrate the possibility of new kinds of phenomena, such as might be produced in artificial agents, other biological systems, or even in humans with special training, new social conditions, brain damage, mental disturbance, etc. This helps us to clarify our concepts by forcing us to define how they apply to these previously unconsidered cases.

## 4. Notes on the design-based approach

There are several variants of the design-based approach. This section describes some of them and attempts to deal with some potential misunderstandings of the preceding description.

### *4.1. Actual vs ideal design-based work*

Very little work in AI has so far explicitly addressed all of (a) to (e), except perhaps in connection with small models of tiny fragments of an intelligent system, though even there it is more usual to present only one design, which is rarely analysed in depth. In the worst cases all that happens is that the program is presented as a 'theory' and shown to pass some tests — what John McCarthy once called the "Look ma: no hands!" approach. Outside AI, control engineers who model control systems, e.g. using sets of differential equations, come closest to meeting the ideal sketched here, though usually they need not simultaneously address as many different levels of virtual machines as would be required for intelligent systems, nor would their methods work for systems involving structure-manipulation and a developing architecture. Biologists often conform (approximately) to the design-based approach, e.g. when they study how the physical and chemical structure of some portion of a cell enables it to do a particular job required for the functioning of that cell, and also compare it with slightly different structures in other cells in the same or a different organism. Another example is analysis of how different flower designs meet different requirements for reproduction.

Most of the design-based work in cognitive science and AI hitherto has been concerned with relatively small fragments of intelligent agents, e.g. parsing, low level vision, planning, learning, etc., all restricted to small subclasses of phenomena. Examples of more ambitious yet very sketchy design-based theories can be found in Minsky [10], Johnson-Laird [7], and chapters 6 to 10 of Sloman [16], with further elaboration in [19] and [21]. Work on Soar [12] appears to have several of these levels, though still mostly with a rather narrow focus on problem solving activities.

### *4.2. Design does not have to be top-down*

The use of the word "design" may suggest to some readers that this approach is concerned only with *top-down* studies that follow the traditional formal design methodology of certain software engineers, as if (a), (b) and (c) had to be sequential stages. This is not the case. As in engineering, it is possible to try working top-down from high level specifications to detailed implementations, or bottom-up by combining known mechanisms to see what sorts of systems can be built from them (e.g. Braitenberg [2], Brooks [3]). More common is a mixture of top-down, bottom-up and middle-out analysis.

A limitation of pure top-down synthesis of very complex designs is that the search space is horrendously complex, so that we risk exploring design-space forever without producing anything that works as required and which can be implemented on mechanisms with the required properties (e.g. brain-like mechanisms), though the search can be reduced if guided by studies of existing systems. A similar limitation confronts entirely bottom-up approaches that build incrementally on simple, well-understood, mechanisms: the search space (of possible designs) is so large that without some guidance from top-down analysis the bottom-up explorations may endlessly roam around designs mimicking primitive organisms that lack the properties we are trying to explain or replicate. Combining the two approaches provides a greater chance of success.

Another natural misunderstanding would be the assumption that all the designs studied should be produced by human beings. This is not so: some people are exploring automated design processes, including mimicking biological evolution by allowing genetic algorithms to produce designs, though it is not clear where the evaluation functions for such experiments should come from: if survival were the only criterion then such experiments might end up with nothing but designs for insects, which are among the most successful survivors.

### *4.3. Variations within the design-based approach*

The distinction between top-down and bottom-up strategies is only one among several dimensions in which design-based studies of intelligence vary. Another dimension concerns whether the study is restricted to the *human* mind or includes other types of systems, such as other animals.

Even among those who study the human mind some are concerned only with attempting to understand a rather narrow class of 'intellectual' capabilities, whereas others wish to encompass a broader range of phenomena, including sources of motivation, personality, emotions, moods and the like.

Some who go beyond the study of human beings restrict themselves to *biological* systems (e.g. modelling capabilities of monkeys, cats, spiders), whereas others have a grander vision of a more general science studying the space of designs not only for biological systems that actually evolved but others that might have evolved, and also possible designs for totally artificial autonomous agents that use quite different basic building blocks from biological systems and are therefore subject to different implementation constraints (e.g. Sloman [17]).

Yet another dimension along which researchers vary concerns whether they restrict themselves to considering only *computational* implementations, or whether they also consider non-computational mechanisms, such as chemical processes. Of course, insofar as the computational/non-computational distinction is somewhat blurred, this dimension is unclear, and it is therefore silly to argue over whether AI requires only *computational* mechanisms, which is why philosophical debates about 'Strong AI' are of little relevance to AI as science or engineering (compare [20]).

Among those who restrict themselves to computational mechanisms there are differences according to whether they are committed to a particular approach or not, e.g. using logic, using connectionist networks, using AI symbol-manipulation languages. In our current state of ignorance concerning which designs are possible and what their properties are, it is pointless to legislate on these matters: if we force everyone to use the same techniques we risk missing important discoveries. Some of these disputes are concerned with formalisms for use at 'compile-time', for generating the agent, others with notations required by the agent itself, at 'run-time'. (I argue in [15] and [18] that intelligent agents need *multiple* forms of representation. Compare Hayes [6]). Some formalisations, such as logics of belief or desire, are more concerned with *external* descriptions of an intelligent agent, and are therefore mainly relevant to specifying requirements, of type(a).

Some who follow the design approach find it useful to build *simulated* environments and *simulated* agents operating in those environments. Others claim that this is 'cheating' and insist on having *real* robots in a real physical environment, for example. Of course, this is a silly dispute as long as those building simulations make their assumptions clear and do not over-generalise their conclusions. This is no different in principle from the role of thought experiments in physics where the implications of extremely idealised situations are analysed in order to clarify problems, concepts and theories.

The simulated/real conflict often goes with a distinction between people who work only on *fragmentary* portions of an intelligent agent (vision, language, planning, etc.) and those who are interested in designing *complete* systems. However, criteria for completeness can also vary: some who build ‘complete’ robots do not attempt to give them autonomous motivational systems, for example, nor human curiosity, imagination, aesthetic preferences, etc. and in that case they are nothing like complete models of human capabilities, even if they are complete robots.

That contrast tends to be related to the long term goals for which the research is pursued: some people do AI essentially as an engineering exercise, and are interested in human capabilities only insofar as these either provide useful ideas, or help to define the environment in which machines will have to work (i.e. the machines may have to interact with humans). Others are more interested (in the long term) in finding scientific explanations for capabilities of existing intelligent agents.

I have tried to show that what I’ve called the ‘design-based’ approach covers a wide variety of alternative and complementary research strategies. Unfortunately, the proponents of one approach will often disparage people who follow others. This is usually due to a lack of insight into the limitations of their own approach and a failure to understand the need to combine different approaches in order to understand the most complex mechanisms known to science.

## 5. Putting it all together

An autonomous human-like agent needs to include components with all the different sorts of capabilities that have hitherto been studied in AI, and more besides. In particular, anything like a human being or other animal will not be driven by some single high level goal, but will have a host of different motivational mechanisms all coexisting and competing with one another, some concerned with bodily requirements others with intellectual or social needs, some short term others long term, some self-centred others related to the needs of other individuals or even the whole of mankind, or other threatened species. (Compare Minsky [10].)

The computational study of human motivation and related phenomena such as emotions, moods and the like is still in its infancy. We don’t yet even know which of the things that seem to be independent causal factors have their own mechanisms and which are merely ‘emergent’ phenomena, arising out of complex interactions between sub-systems whose real function is quite different. Some of the early studies in this area built models in which states like anger, fear, joy were represented by explicit variables, or database-entries, which could have numeric values, or symbolic values like “low”, “medium”, “high” etc. My guess is that these designs are totally misguided and that those affective states are best thought of as emergent properties of processes that serve more specific purposes in the total system.

Of course, insofar as an intelligent agent includes some internal self-monitoring capability it may learn to detect the patterns involved in such global emergent states, and as a result there may be an explicit internal representation of the state (as when a person not only *is* angry but also *feels* angry), but that kind of self-monitoring is a secondary process, not an essential part of the original state, which can often occur without being recognized by the agent.

However, these are issues on which it is too early to be dogmatic: we need to explore alternative approaches, and see what we learn.

## 6. The structure of design space

It is worth commenting on the 'shape' of design space. People who think they understand what they mean by intelligence or consciousness often assume that such concepts correspond to dichotomies, which divide things in the world up into two mutually exclusive classes: the entities that are instances of the concept (e.g. have minds, or are conscious) and those that are not. This is easily undermined by showing the implausibility of finding a clear division among living organisms which are and which are not conscious, or intelligent.

That often leads to another mistake: the assumption that the space is a continuum, with smooth variation and no clear point at which any important distinction relevant to the nature of mind can be made. This is a mistake because anyone who has explored real design options is aware that there are many discontinuities. A simple example is the difference between two programs, one of which has an extra 'elseif' clause in a multi-branch conditional. There is no possibility of half of such a clause, or a quarter, etc.: the clause is either there or not there. Changes in physical network topologies, or in abstract network structures, are also discontinuous. Some of these design discontinuities make important qualitative differences to the overall capabilities of the system, whereas others (e.g. adding an extra memory location) may make only a marginal difference, unless a threshold is exceeded.

One of the tasks on which work has barely begun is to explore the discontinuities in design space to see what different kinds there are, and what difference they make to the capabilities of an agent or organism. Such a study may help us understand better the evolution of different sorts of intelligent capabilities in different animals.

## 7. Conclusion

Readers wanting more detailed illustrations of the design based approach will find many in this volume, some prospective some retrospective.

This discussion of prospects for AI has not mentioned arguments by Searle and others purporting to show that AI cannot possibly provide an explanation of anything like human intelligence because human beings have capabilities which can be shown to be impossible for computational systems. Such arguments, some based on Gödel's incompleteness theorem, some based on philosophical analysis, some based on thought experiments in which implications of the success of AI are explored, all purport to refute the so-called "Strong AI" thesis. I shall not attempt to answer these arguments here because I have already done so in [20], where I try to show, among other things, that there are at least eight distinct interpretations of the Strong AI thesis, on some of which it is patently false and not worth attacking, whereas others pose interesting and open questions, which can be answered only by further investigation of design possibilities, not armchair pontification.

There's lots more to be done. Hubert Dreyfus, the arch-critic of AI, once likened AI to trying to get to the moon by climbing trees. Such comments ignore the fact that pathways to scientific knowledge and achievement have never been simple and straight. Perhaps early tree climbers were going through essential steps towards understanding the physical universe, without which space travel would never have been achieved by their descendants?

We have already learnt a great deal about the nature of the problems, and why some tempting mechanisms don't work and why others are better. Even if few specific notations, techniques, algorithms or architectures from twentieth century AI survive in explanatory theories two centuries hence, it could still turn out that our current explorations were an essential part of a profound learning process. Germs and seeds don't have to look like what they grow into.

**References**


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

[2] Braitenberg, V. (1984) *Vehicles: Experiments in Synthetic Psychology* Cambridge, Mass: The MIT Press

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

[4] Clore, G.L., Ortony, A. (1988): 'The Semantics of the affective lexicon' in V. Hamilton (ed) *Cognitive Perspectives on Emotion and Motivation,* Kluwer Academic Publishers 367–397

[5] Grossberg, S, (1987) 'Competitive learning: from interactive activation to adaptive resonance', *Cognitive Science*, 11, 23-63

[6] Hayes, P.J. (1974) 'Some problems and non-problems in representation theory', in *Proceedings AISB Summer Conference* University of Sussex, 1974 (out of print), reprinted in *Readings in Knowledge Representation*, eds. R.J. Brachman and H.J. Levesque, Morgan Kaufmann, Los Altos, 1985, 4-21.

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

[8] Johnson-Laird, P. N., Oatley K.(1989) 'The language of emotions: an analysis of a semantic field' *Cognition and Emotion* **3**,2 81–123

[9] Marr, D. (1982) *Vision*, W.H. Freeman, San Francisco

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

[11] Ortony, A., Clore, G.L., Foss, M.A. (1987) 'The referential structure of the affective lexicon' *Cognitive Science,* **11** 341–364

[12] Rosenbloom, P.S, Laird, J.E., Newell, A and McCarl, R. (1991) 'A preliminary analysis of the Soar architecture as a basis for general intelligence', *Artificial Intelligence,* **47**, 289-325.

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

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

[15] Sloman, A. (1975) 'Obituary notice: Analogical Representations', *AISB Quarterly* 1975, re-printed as 'Afterthoughts on analogical representations' in *Readings in Knowledge Representation*, eds. R.J. Brachman and H.J. Levesque, Morgan Kaufmann, Los Altos, 1985, 432-439.

[16] Sloman, A. (1978) *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind* Harvester Press, and Humanities Press. (Now out of print.)

[17] Sloman, A. (1984) 'The structure of the space of possible minds', in *The Mind and the Machine: philosophical aspects of Artificial Intelligence,* ed. S. Torrance, Ellis Horwood,

[18] Sloman, A. (1985) 'Why we need many knowledge representation formalisms', in *Research and Development in Expert Systems* ed. M Bramer, Cambridge University Press, 163-183,

[19] 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

[20] Sloman, A (1992) '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, 355-396

[21] Sloman, A (1992) 'Prolegomena to a theory of communication and affect' in Ortony, A., Slack, J., and Stock, O. (Eds.) *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, Heidelberg, Germany: Springer, 1992/3, 229-260. (Also available as Cognitive Science Research Paper, School of Computer Science, The University of Birmingham.)