Explorations in Design Space

Aaron Sloman and the Cognition and Affect Group [1]

Abstract. The methodology of AI as the general study of natural and artificial intelligence is outlined, including exploration of designs for a variety of behaving systems, for both scientific and engineering purposes. The 'design-stance' treats minds as sophisticated self-monitoring, self-modifying control systems. Different architectures can satisfy that description. Exploration of possible *requirements* and *designs* for complete agents is important not only for engineering purposes but also for bringing together hitherto fragmentary studies of mind in various disciplines, providing a basis for an adequate set of descriptive concepts, clarifying various evolutionary questions, and making it possible to understand what goes wrong in various human activities and which remedies might work. This requires studying niches as well as designs.

1 INTRODUCTION

Few people appreciate the breadth of AI. Many construe it as merely a branch of engineering. Some construe it as restricted to 'central' cognitive processes, such as planning and inference. Some restrict it to rule-based symbol manipulations. Most are unaware that AI as actually practised and reported is a broad unifying discipline – the general study of sophisticated self-modifying control systems, whether natural or artificial, whether studied for practical or scientific purposes. From this viewpoint, AI is the exploration of the space of possible designs for complete agents. [2]

Much AI work understandably addresses restricted capabilities (e.g. perception, planning, etc.), using specific mechanisms (e.g. procedural languages, rule-based systems, neural nets, etc.) and forms of representation (e.g. logic, rules, images, connection weights). We have yet to learn how to combine the fragments. This requires a framework within which many different sub-capabilities can be integrated, i.e. an architecture, possibly a *dynamically changing architecture.*

Studying architectures has both *scientific* objectives concerned with understanding existing agents, and also *practical* objectives discussed below. However, the task lacks definition, partly because we have no general theory of types of architectures, and partly because we lack clear and unambiguous specifications of *what* is to be explained, or modelled. For example it is notoriously difficult to reach a consensus on how to define notions like 'consciousness', 'perception', 'learning' and 'emotion'. Many such terms are used with different meanings by different groups of scientists, causing spurious disagreements. Explorations in design space will identify more precise possible interpretations ([8], [9], [15]).

Many who study whole systems (some philosophers, psychologists, ethologists, etc.) have no idea how to think about underlying explanatory mechanisms: they are trained only to observe global behaviour and look for 'external' regularities, or to describe the introspective phenomenology, even though behavioural and phenomenological descriptions grossly underdetermine the underlying explanations. They sometimes use loose metaphorical ways of posing alternatives (e.g. 'Was the subject thinking in words, or in images?'), without realising that each 'explanation' is compatible with significantly different implementations with very different consequences. Attempts at implementation expose such ambiguities.

Some people aim for neural explanations: but not enough is known about how neurons actually work, and there is a risk that neuroscience will tell us only about low level organisation, without addressing important higher level functions, just as studying physical computer components tells you nothing about the operating system, nor which high level algorithms or data-structures are employed. Looking for empirical correlations between high level processes and physical or physiological changes may fail, as explained in section 6.

2 PUTTING AI TOGETHER

AI work on complete agents often targets narrow practical problems, e.g. designing a welding robot. There are some models of complete agents, but usually they are very restricted, e.g. robots without motivational autonomy, or with limited learning powers restricted to modifications within a module, and very rarely with an architecture that develops over time, producing new sub-mechanisms and new links between mechanisms, as humans seem to do, e.g. learning new forms of self-control.

There are attempts to design more general human-like agent architectures (e.g. [5], [7], [14]), but they are still very sketchy. E.g. a full account of human perception would include a comprehensive survey of perceivable aspects of the environment ([4], [10]) and different modes of perception (noticing, searching, controlling, etc.); and a full theory of learning would have to cover such diverse phenomena as infant development, acquisition of sporting skills, university-level learning, and re-

---

[1] School of Computer Science & Cognitive Science Research Centre, The University of Birmingham, Birmingham B15 2TT, UK.
Email: A.Sloman@cs.bham.ac.uk
The group includes Luc Beaudoin, Glyn Humphreys, Christian Paterson, Edmund Shing, Tim Read, Ian Wright. Our work is supported by the Renaissance Trust and the UK Joint Council Initiative (Grant Ref MRC SPG9200393).

[2] This is one of several papers presenting the 'design-stance' view of a mind as essentially a sophisticated self-monitoring, self-modifying control system: [Sloman & Croucher 1981, Sloman 1989, 1992a/b, 1993a/b, 1994a/b] [Beaudoin & Sloman 1993] [Beaudoin 1994].

ECAI 94. *11th European Conference on Artificial Intelligence* Edited by A. Cohn
Published in 1994 by John Wiley & Sons, Ltd.

covery from brain damage.

Moreover, a single architecture gives a shallow explanation, whereas showing in what ways it is better or worse than other designs gives a better understanding and, by revealing trade-offs, helps us to understand evolutionary pressures. Comparing and analysing alternative sets of requirements and designs I call 'exploration of design-space'. This serves engineering as well as science ([13]).

3 PRACTICAL IMPLICATIONS

Building useful machines is not the only practical reason for studying architectures in AI. Another is trying to help when something goes wrong with a complex *natural* architecture. Where problems involve interactions between different components, it is desirable to have some idea of the architecture in order to be able help improve things in a reliable way. E.g although 'craft' remedies based on intuitive skills may work in limited situations, ignorance of how they work means that we cannot predict boundaries of applicability.

Improved understanding of mental architectures could lead to improvements in educational procedures, various kinds of therapy, counselling, HCI design, social engineering and management. Good *practical* solutions need to be based on good *explanatory* theories. These enable us to deduce conditions under which different therapies can help. Cognitive-behavioural therapy, which tries to change patients' incorrect beliefs, may help in some cases, while other problems have different causes, e.g. wrong or missing control links, or a lack of certain kinds of internal self-monitoring. Different flaws need to be repaired in different ways. Some may require re-structuring of the architecture, for instance like training in sport, music or acting that develops new reflex links, or new forms of control.

Some familiar system descriptions, e.g. 'thrashing', 'load average', 'interrupts', 'memory fragmentation', etc. can be understood *only* in terms of interactions within an underlying architecture. Less obviously, familiar labels for mental states (e.g. 'believes', 'understands', 'perceives', 'desires', 'emotion') also presuppose very complex architectures. Exposing such presuppositions is hard without an account of interactions between sub-mechanisms and their roles in the whole system. E.g. colloquially 'attention' (sometimes) seems to refer to processes that make selections: for instance between perceptual mechanisms (e.g. vision *vs* touch), between stored information items, between modes of processing of information, between goals, between plans, etc. A theory of the architecture will show which kinds of selections are possible, or necessary.

Some emotions involve reduced control of attention ('perturbances'): e.g. grief makes it hard not to dwell on lost loved ones ([17] [11], [3]). A theory of the underlying architecture could show what sort of control is involved and what causes the interference. Ignorance about attention control mechanisms could impair therapy, counselling or education.

Besides doing research that deepens our understanding, we also need to find good ways to *teach* the new concepts and theories. The teaching of psychology, and training of teachers, therapists, social workers, and managers might all benefit. In particular, transparent and modifiable implementations of simplified architectures could be powerful devices for teaching new ways of thinking about mental processes.

4 ANALYSING REQUIREMENTS

Merely producing a good design and implementation are never enough. We must understand which set of requirements is satisfied and *why*, and, if possible, how each solution compares with alternatives. Requirements may be explicit in an engineering project or implicit in biological evolution, corresponding to an ecological niche (though, as pianists demonstrate, an organism's capabilities may exceed the requirements of its species' niche). The physical environment does not determine the nice: the niches for butterflies and a monkeys in the same part of a forest are totally different. Sometimes one part of a design determines the requirements for another. Two organisms may feed off the same plant, one biting off large chunks of the plant while the other delicately selects seeds from pods: their perceptual requirements will then be very different. They inhabit different environments in the same place!

Analysing design requirements for agents as complex and multi-faceted as human beings is a complex task. It is easy to oversimplify requirements, like the brain scientist who once said to me that the function of vision is simply to inform about distances to contact. Likewise, grossly oversimplified definitions of 'motivation', 'learning', 'emotion', etc. can misdirect the search for explanations. Many philosophers wrongly assume that rationality is a requirement: intelligent behaviour is often based not on rational principles, but, for instance, on learnt heuristics activated by pattern matching under time pressure. A rational design can produce irrational behaviour!

As well as alternative *designs*, a general study must survey alternative sets of *requirements* (niches) and the topology of the space of such niches. Evolution has explored many anatomical and physiological architectures matching different sets of requirements. There is probably a similar variety of control system architectures, meeting diverse requirements, though these are less open to observation and leave no fossil records. There is no agreed language for specifying behavioural requirements, or niches, including relevant details of 'affordances' in the environment ([4]). Many concepts used by scientists for describing the environment are either too impoverished, e.g. numerical measures, or too ill-defined, e.g. ordinary language. Formal grammars, structural descriptions, and the like, can help; including possibly grammars for forms of internal or external behaviour.

Behavioural requirements for an orang-utang are very different from those for a dolphin, though there is some overlap. Various constraints on designs must be considered: e.g. evolutionary constraints, physical environment, social structures, the degree of dependence of neonates, how fast environments and their affordances change, available processing mechanisms. There is no unique concept of an intelligent system, and no unique set of requirements for intelligence: different kinds of control systems have different clusters of capabilities. Designs involve trade-offs and compromises, and there is no general optimality measure, so we must not expect unique mappings between niche space and design space.

5 USEFUL SIMPLIFICATIONS

Galileo, Newton, Einstein and others used key simplifications to produce major scientific advances. The daunting tasks of specifying requirements for human-like architectures and exploring adequate designs can be simplified in various ways, including:

- Ignoring some design constraints temporarily, e.g. ignoring neural implementability, or evolutionary plausibility,
- Ignoring some requirements, e.g. social requirements.
- Focusing on isolated components, e.g. vision, learning, NLP.
- Building in 'pre-compiled' competence rather than the ability to learn or develop the abilities.
- Focusing on complete but very simple agents with relatively few capabilities, e.g. primitive biological organisms.
- Exploring designs with a wider variety of functions, but with shallow implementations for those functions: Bates et al. [1] call this a 'broad and shallow' architecture.
- Simplifying the environment and the agent's interface to it, making perception and action easier to implement.
- Focusing only on low level architectural issues, e.g. designing neural circuitry rather than the high level motivation, planning, etc.
- Ignoring the individual variability in humans.
- Ignoring variability between species.

Though easy to criticise, such simplifications (including 'toy worlds') can advance understanding, if done in a disciplined fashion without inflated claims. Design decisions not derived from requirements can be justified for research purposes in various ways, e.g. by consistency with previous decisions, by empirical evidence concerning how people do things, by project resource constraints, by the lack of any clear alternative, by the goal of exploring their consequences, etc. Unfortunately, people who make different sorts of simplifications are often combative and over-critical of one another, and approaches that complement one another are often presented as rivals. Apparent disagreements may merely reflect an interest in different levels (e.g. neural *vs* cognitive).

Of course, *oversimplifying* can compromise relevance, e.g. representing processes as changes in a fixed set of numerical parameters may be a mistake if coping with internal and external *structural* change is a major determinant of the system being modelled. Defining vision as detection and description of structure and motion ([6]) could also be a serious mistake, if it leads researchers to ignore other 'affordances' ([4] [10]).

Design studies need not all go top-down from requirements via designs to implementations: some explorations are 'bottom-up' or 'middle-out'. Designs and implementations can reveal inconsistencies in requirements, leading to revisions. Experimental implementations often suggest both additional requirements and new design possibilities; and flaws in performance can expose missing requirements and thereby design constraints. More generally, architectures for complex behaving systems can be studied in several different ways:

- They can be studied top-down (from requirements to designs) or bottom-up (i.e. exploring primitive mechanisms and seeing how they can be made more complex, or combined in different ways), or middle out.
- They can be studied in their own right or as things that evolved from earlier designs.
- They can be broad and deep, broad and shallow, narrow and deep or narrow and shallow!
- They can be studied at the physical level or at higher levels of abstraction as virtual machines possibly at different levels in an implementation hierarchy.
- Links between levels of virtual machines may be *implicit* in computer demonstrations, or, preferably, explicitly *derived.*

6 SUPERVENIENT MACHINES

Some architectures are non-physical. Newell and Simon emphasised *physical* symbol systems whereas symbols in *virtual* machines are often more important, e.g. when very rapid structural reorganisation is required (cf. [15]). 'Hard wiring' precludes some types of self-modifiability. Every programming language defines a virtual machine: e.g. for manipulating numbers, strings, lists, logical clauses, etc. These are *implemented* in physical machines. Sometimes there are several layers of virtual machines. Implementation mechanisms can be very complicated but they are not mysterious – being non-physical does not imply being mystical in any way. In philosopher's jargon virtual machines 'supervene' on physical ones. Sometimes part of the implementation is in the environment: e.g. semantic states that refer to particulars or relationships in the environment ([15]).

Implementation or supervenience need not use simple correspondence of parts or sub-processes. Sparse virtual arrays may have far more parts than the underlying physical machine; and processes involving them can change very many array components whilst changing few physical components. Similarly, if one database uses deduction to *derive* items that are *stored explicitly* by another, their underlying physical structures and processes will be very different although they are functionally equivalent.

Relationships between virtual and physical machines may be many-many: Two very different operating system architectures can run on the very same underlying 'physiology', and the same operating system can run on two physically different machines. Similarly people with very different languages, cultures and personalities may have physically similar brains. So studying physiology may not give information about important virtual machines. Some brain structures may indicate only generic low level implementation details, telling us nothing about important virtual machine characteristics. Thus the philosophical assumption that supervenience implies correlations between mental states and brain states is wrong. Besides, mere correlations don't explain: we need to know *how* one thing is implemented in another so as to provide a *derivation* that the implementation is adequate – like proving that a computing implementation meets its requirements ([13]).

7 TERMINOLOGICAL CONFUSIONS

Specifying what we are trying to model or explain is hard when people often disagree on terminology and argue endlessly at cross-purposes ([8]). E.g. there are many significantly different definitions of 'emotion': in terms of external eliciting conditions, behavioural responses, proprioceptive feedback, cognitive contents, neural processes, and others. 'Intelligence' and 'consciousness' are almost as bad. Deeper theories about the underlying mechanisms will help us define precisely what kinds of phenomena we are talking about, just as discovering the architecture of matter led to good theory-based definitions for concepts of types of stuff, e.g. 'water' and 'salt'. The theory generated a much richer set of concepts of kinds of stuff than had previously been dreamed of. I expect the same to happen for mental concepts: in deep science good definitions follow good explanations.

Just as understanding how operating systems work enables us to grasp concepts describing their possible states, such as 'thrashing', so a theory of an architecture capable of supporting human-like abilities will help us classify mental states and processes, generating a set of concepts refining and extending our mental vocabulary. Concepts like 'attention', 'consciousness', 'belief', 'image', 'intention', 'emotion', 'mood' will evolve into or be replaced by more precisely defined notions, including distinctions between cases that we now group together, and concepts for possibilities never yet encountered. We won't find 'correct' definitions: some existing labels, e.g. 'emotion', 'attention', will evolve into families of labels for importantly different cases (such as 'perturbance', mentioned above.)

This should lead to a deeper understanding of possible pathologies, and make it easier to design successful therapies. Unfortunately, our understanding of which architectures are possible, what their behavioural implications are, what their physiological requirements are, and how they can go wrong, is still primitive. Perhaps we'll one day discover virtual machines which differ from what we now understand as much as modern operating systems differ from early mechanical calculators. Such discoveries will underpin new concepts of structure, state and process.

8 DESCRIBING ARCHITECTURES

Describing architectures in a precise and productive way requires a language for specifying requirements, describing functional decomposition, and describing states and behaviour of both complete systems and their components. At present our vocabulary is limited and confused: we need a calculus of causation, including, but going beyond, simple-minded notions like 'cause', 'produce', 'prevent', 'preserve', 'protect', 'monitor', 'modify', 'accelerate', etc. Control engineers and computer scientists have extended this vocabulary e.g. by talking about positive and negative feedback, amplification, damping, filtering, interrupt handling, procedure calls, variable binding, rule compilation, exception trapping, and so on.

Higher level control concepts are needed for analysing mental processes, and talking about interactions within relevant virtual machines. Concepts like 'algorithm' and 'proof' have received much attention, but don't help us think about architectures ([12]). AI programming languages and systems constitute one kind of attempt to define appropriate virtual machine concepts (e.g. SOAR and neural nets). But we don't yet know what sorts of virtual machine can support architectures with sufficiently rich explanatory potential.

Architecture-based concepts should figure in courses in psychology, philosophy of mind, psychiatry or education. This does not imply that ordinary mental concepts will be *eliminated* in favour of lower level concepts like neural state, or as some have suggested, concepts of the phases of the state space of a complex dynamical system. Ordinary concepts are too useful to be completely wrong, so they are likely to be refined and extended, not rejected ([14] [15]). Good theories will not necessarily exactly support familiar notions like 'belief', 'desire', 'emotion', 'learning', etc. Rather, the newly generated concepts will be partly similar to the old ones and partly different, as happened with concepts of physical stuff.

Our language for describing causal interactions needs more specific terms for high level virtual machine structures, mechanisms and types of processes. We need concepts both for control functions and also for information-bearing control substates. [3]The Birmingham *Cognition and Affect* project (footnote 2) uses notions like 'management' and 'meta-management' processes, 'asynchronous goal-generactivators', 'attention filters' with 'interrupt thresholds', 'insistence' and 'intensity' of motivators, and hierarchies of control dispositions corresponding roughly to personality, attitudes, moods, and 'perturbances', mentioned above, though our concepts are not yet sufficiently general, precise, or comprehensive([2], [3]).

9 EXPLORING DESIGN SPACE

Biologists often compare and contrast related plant or animal structures and their functions. Comparative studies are essential for a deep understanding: we need to explore the structure of design space, and not simply look for one architecture (or type of architecture) to 'explain intelligence'. Fully understanding an architecture includes knowing the functional roles and causal links of major components and how they relate to specific requirements. This means knowing what difference the presence or absence of those components or links would make to capabilities of the total system. Such comparative explorations could produce both a hierarchical classification of *requirement sets* (niches) and a separate taxonomy of *designs*, with mappings between the two. (As already explained, the mappings are not one-to-one.)

A generic design accommodates a wide variety of individual differences resulting from slightly different 'initial parameters' or from environmental differences. A particular self-modifying architecture for infants may permit enormously wide variation in subsequent development, while imposing some general limits. Improved understanding on this point should settle nature-nurture debates. We need to study both individual differences and also differences across larger design classes, e.g. human beings *vs* chimpanzees, primates *vs* other mammals. Comparing close neighbours in design space helps us understand the features in which they differ. Looking further afield reveals the significance of design features shared only by close neighbours. E.g. we can better understand the role of architectural characteristics that we share with chimpanzees if we also study organisms that lack them.

Good languages for specifying requirements, designs, and relationships between them, will generate taxonomies of niches and architecture types, including many discontinuities, such as differences in: numbers and types of components; numbers of levels of virtual machines in the implementation; kinds of coupling between low level and high level virtual machines; whether the management structure is roughly hierarchical, or anarchic, or 'voting'-based as in some neural models; which kinds of information store are available; which forms of representation and representation-manipulating mechanisms are

---

[3] When describing how to get from the kitchen to the front door I may use a 'mental image' of my house. But it is not clear what such talk about images amounts to. Is the image an object that causes something to happen, or is that just a misleading manner of speaking? What about knowing how to ride a bicycle, or how to discern a composer's style? There are many similar unanswered questions

available (e.g. logical capabilities); whether memory is garbage collectable; whether processing is mainly serial or whether concurrent activities are supported (Why is human concurrency at high levels so limited?) and many more. We don't know how many important dimensions of variation of architectures there are. (Cf. [14])

10 CONCLUSION

AI needs a high level vision of its goals and methods. I offer exploration of design space and its relationships to niche space as such a vision.

The design-based study of architectures for intelligent (or nearly-intelligent) agents is important not only for engineering purposes but also for bringing together hitherto fragmentary studies of mind in various disciplines, for providing a basis for an adequate set of descriptive concepts, and eventually for making it possible to understand how various human capabilities can go wrong and which remedies are most likely to work.

There is no one 'right' way to do the research. It is less likely to lead into blind-alleys if people using bottom up and top down approaches communicate and cooperate, instead of each claiming to have the only correct methodology. Biology-based and engineering-based studies can also inform one another. Different but complementary explorations should eventually converge on an multi-level but integrated explanatory framework and a new common language.

As the exploration progresses we can expect it to yield several new systematically generated families of concepts, for instance a hierarchical family of concepts corresponding to different sets of requirements for agents ('niche space'); a different family of high level (virtual machine) design specifications ('design space'); a family of implementation mechanisms; and for each design or design family a collection of concepts corresponding to possible states and processes supported by that design.

I have left open whether the resulting designs will use only computational mechanisms, mainly because I think it is a relatively unimportant question, for reasons given in [12] and [14]. Some control systems use analog mechanisms that can be approximated though not replicated by digital computers. If non-computational mechanisms are needed for embodied agents, AI will expand to include them (e.g. servo-mechanisms in robotics.)

There is a vast amount of extremely difficult work still to be done. It has been claimed that AI is dead. 'Dead' is hardly an accurate description for a young discipline that still has many important unfinished tasks that no other seems able to tackle. But major advances may take hundreds rather than tens of years.

SHORT BIBLIOGRAPHY

[1] 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.).

[2] L.P.Beaudoin & A.Sloman (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

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

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

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

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

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

[8] 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.

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

[10] 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

[11] 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, The University of Birmingham.)

[12] 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, Birmingham University)

[13] 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.

[14] 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

[15] Sloman, A (1994a to appear) Semantics in an intelligent control system, in *Proc. British Academy and Royal Society Conference: Artificial Intelligence and The Mind: New Breakthroughs Or Dead Ends?*

[16] Sloman, A. (1994b) (in preparation), Representations as control sub-states,

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