INCOMPLETE DRAFT: September 14, 2008

The Cognition and Affect Project: Architectures, Architecture-Schemas, And The New Science of Mind.[1]

Aaron Sloman
http://www.cs.bham.ac.uk/˜axs/
School of Computer Science
University of Birmingham, UK

NOTE

This is a slightly updated version of the report to John Salasin and Robert Rolfe, produced in July 2003 (see note 1 below), and made available online as

http://www.cs.bham.ac.uk/research/projects/cogaff/misc/darpa-architectures-draft.pdf

This revised version was produced in October 2004, but not placed on the web site.

However even this version is now out of date, having been superseded by this report on the CogAff Project (which may be updated from time to time):

http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200307
Progress report on the Cognition and Affect project:
Architectures, Architecture-Schemas, And The New Science of Mind
(Original 2003. Revised October 2004 and 2008.)

[1] The first draft was written in response to a request from John Salasin and Robert Rolfe for their DARPA Integrated Cognition Project, in 2003. Since then this has been revised and extended.

Abstract

Research on algorithms and representations once dominated AI. Recently the importance of architectures has been acknowledged, but researchers have different objectives, presuppositions and conceptual frameworks, and this can lead to confused terminology, argumentation at cross purposes, re-invention of wheels and fragmentation of the research. We propose a methodological framework: develop a general representation of a wide class of architectures within which different architectures can be compared and contrasted. This should facilitate communication and integration across sub-fields of and approaches to AI, as well as providing a framework for evaluating alternative architectures. As a first-draft example we present the CogAff architecture schema, and show how it provides a useful framework for comparing and contrasting a wide range of architectures, including H-Cogaff, a proposed architecture for human-like systems. All of these concern *virtual machine architectures* whose natural implementations use biological mechanisms but some of which may use products of human engineering.

Besides attempting to understand what sorts of virtual machine architectures are possible we also need to answer a prior, perhaps deeper, question: what are the architectures needed for? This has two aspects: (a) if we are attempting to model natural intelligence, whether in humans or other organisms (e.g. nest-building birds) then we need to find out what the capabilities of those organisms are: and that may involve learning to look at them in new ways, whereas (b) if we are trying to build systems for practical applications then we need to be very careful to identify the requirements of those applications. Merely describing something as “human-like” for instance does not specify what capabilities it has.

Whether the task is of type (a) or type (b) it requires us to have a good characterisation of two spaces and their relationships, namely *design space*, the space of possible designs and *niche space*, the space of possible sets of requirements. In particular we need to understand niche space in order to know what questions to ask, and we need to understand design space in order to know what possible answers there are. We need to understand mappings between the two spaces in order to understand how to compare different answers to the same question. And we need to understand trajectories in those spaces in order to understand evolution, development and learning. If we ask the wrong questions, e.g. because we mischaracterise niche space, then we may solve some easy problems without even noticing the far more difficult, still-unsolved problems.

Contents

1 Background

This is a partial progress report on the Cognition and Affect project. It all started when the late Max Clowes gradually persuaded me around 1970 (when I was at the University of Sussex) that the best way to make progress in various fields of philosophy (including epistemology, metaphysics, philosophy of mind, philosophy of language, philosophy of mathematics, and philosophy of science) was to understand how to design a working mind, or at least fragments of working minds, since doing it all at once was too difficult. Some of the early results of that 'conversion' were reported in 1978 in *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind*.[2] It was already clear then that understanding architectures was central to the task, and Chapter 6 reported some preliminary explorations which were followed in subsequent chapters with some discussions of mechanisms and structures for learning about numbers, the need for multiple forms of representation, the architecture of a visual system, and other things. The book was, in part, an attempt to explain why the problems were very hard (e.g. see Chapter 9 section 9.13) and why we could not expect rapid progress, contrary to claims made by various leading researchers in AI, prior to that time, and also later.

Moving to a research chair at the University of Birmingham in 1991 (partly) liberated me from undergraduate teaching and enabled me to continue working on the ideas, helped considerably by interactions with members of the psychology department (Glyn Humphreys, Jane Riddoch and colleagues).

1.1 Funding for the research

Partly because of the long term speculative nature of the research, it has proved difficult to get significant funding – most of the grant proposals have been unsuccessful: presumably because the proposals are seen as too broad, too long term, to speculative, and insufficiently focused to fit the normal pattern of a funded project.

However, while at Sussex I was given an unsolicited grant by the GEC Research Laboratory (Chelmsford), which paid for me to go on leave for two years, and the Rennaissance trust (Gerry Martin) also gave me two unsolicited grants, first to support working from home during the GEC fellowship, and later to support a PhD student at Birmingham.

Fortunately, I have had a succession of PhD students for most of whom I did not have to obtain funding, and a few further grants allowing me to appoint research fellows, buy equipment, pay for travel, etc. A grant from the Vice Chancellor's fund paid for Darryl Davis[3] for two years from 1995. An unsolicited grant from DERA Malvern paid for Brian Logan[4] for three years from 1995. A grant from the Leverhulme Trust paid for 36 person-months of effort, paying for Brian Logan (who had helped with the proposal) for a month in 1999, then for Matthias Scheutz[5] (on leave from Notre Dame University), for 14 months from June 2000, and finally Ron Chrisley[6] (on leave from Sussex University) for 21 months from October 2001.

[2] Out of print but now available online, with some recent annotations added: **http://www.cs.bham.ac.uk/research/cogaff/crp/**

[3] **http://www2.dcs.hull.ac.uk/NEAT/dnd/**

[4] **http://www.cs.nott.ac.uk/˜bsl/**

[5] **http://www.cse.nd.edu/˜mscheutz/**

[6] **http://www.cs.bham.ac.uk/˜rlc/** The project ended in June 2003 but collaborative work with the three people concerned continues. The project is described here http://www.cs.bham.ac.uk/˜axs/lev/

The DERA project paying for Brian Logan involved close collaboration with Jeremy Baxter and Richard Hepplewhite at DERA, Malvern, who were using our SimAgent toolkit[7] to develop demonstrations of "Computer generated forces". Darryl Davis, Brian Logan, Jeremy Baxter, Richard Hepplewhite and Matthias Scheutz all made significant contributions to the SimAgent toolkit, in addition to being users. Matthias Scheutz has extended it with his *Simworld* package [REF], and Brian Logan has been doing work on designing and implemented facilities for distributed simulations.[REF]

For a short time after coming to Birmingham I had a studentship from the Renaissance trust, and a studentship with supporting funding (shared with Glyn Humphreys) from the now defunct UK Joint Council Initiative on Cognitive Science and HCI.

In 2003 I was invited, with Jeremy Wyatt, to join a consortium submitting a bid to the EC framework 6 initiative on 'Cognitive Systems'. The bid was successful and a project entitled 'Cognitive Systems for Cognitive Assistants' (**CoSy**) started in September 2004, with 7 partners in total, costing about 7M euros. The Birmingham contribution will build on and extend the work done in the Cognition and Affect project. A summary of the project proposal and plans for work in Birmingham in that project can be found at this web site: http://www.cs.bham.ac.uk/research/projects/cosy/

1.2 Work done by research students

The students working more or less closely on the Cognition and Affect project who have completed are: Luc Beaudoin (who started while I was at Sussex in 1990), Tim Read, Ed Shing, Ian Wright, Christian Paterson, Chris Complin, Steve Allen, Ian Millington (who decided to leave and start up a company[8] before finishing), Catriona Kennedy and Nick Hawes (funded by Sony to develop game agents with 'anytime' planning capabilities)'.[9] Ongoing work is being done by Manuela Viezzer (on ontology discovery), Dean Petters (on attachment in infants), and Dave Gurnell (on collaborative planning). Because I believe that research training includes learning to identify interesting research problems, these students all worked on projects of their own choosing rather than contributing in a closely managed fashion to the overall CogAff project.

Topics investigated by the PhD students have included: architectures for human-like agents, requirements for reinforcement learning mechanisms, planning systems, emotions as emergent phenomena in intelligent systems (including a long joint paper on grief by Wright, Beaudoin and Sloman (Wright et al., 1996)), aspects of evolution (e.g. Chris Complin's PhD on evolution of evolvability), neurally inspired architectures, intelligent self-monitoring software systems, 'anytime' planning in intelligent synthetic game agents, the development of an individual's ontology, the development of different kinds of attachment in infants, and cooperative planning. Several of the students also made important contributions to the SimAgent toolkit: many of its features were inspired by problems they encountered in attempting to investigate architectures.

[7] Described in (Sloman and Logan, 1999) and **http://www.cs.bham.ac.uk/~axs/cogaff/simagent.html**, and reported in this recent overview (Ritter, 2002). The toolkit is available online.

[8] **http://www.mindlathe.com/**

[9] Several completed theses are now online here **http://www.cs.bham.ac.uk/research/cogaff/phd-theses.html**. Paterson's MPhil thesis, and completed PhD theses by Beaudoin, Wright, Complin, Allen, Kennedy and, most recently, Hawes. Source code for Allen's project is also included. Students supervised by Brian Logan at Nottingham University, by Darryl Davis at Hull University and by Matthias Scheutz at Notre Dame University have been doing related work, also using our toolkit.

Catriona Kennedy's work, generalising our notion of *meta-management* to cover *mutual meta-management* (multiple agents observing one another including observing one another's observations), has attracted funding from DSTL Malvern, for work on self-monitoring software systems (Kennedy and Sloman, 2003).

1.3 More general work on the Cognition and Affect project

With the help of the students and the research fellows named, and others, I have also been working on a wide variety of topics connected with the nature of human, animal (non-human), and artificial minds, including the following topics (which subsume the research funded by the Leverhulme Trust on "Evolvable virtual information processing architectures for human-like minds"):

- Showing how unwitting use of 'cluster' concepts[10] causes of confusion in researchers. This confusion is manifested in discussion of many concepts including, for instance, 'emotion', 'intentionality', 'intelligence', 'consciousness' and 'life'. Often the confusion can be reduced by means of architecture-based conceptual analysis: e.g. asking questions about emotions or consciousness in robots or animals may lead to unresolvable disputes if "emotion" and "consciousness" are partly indeterminate cluster concepts. Producing several more refined alternative definitions of each, namely architecture-based definitions, can replace pointless semantic debates with productive research (Sloman, 2001a; Sloman et al., 2005; Sloman and Chrisley, 2003)
- Developing architecture-based analyses of concepts of emotion and other affective states and processes (e.g. mood, desire, pleasure, pain, attitude, intention, preference), and in particular an architecture-based analysis of grief (with Wright and Beaudoin), and architecture-based analysis of varieties of surprise (with Scheutz), attempting to characterise a useful, general architecture-based "affective/non-affective" distinction (with Scheutz and Chrisley).[11]
- Analysis of confusions involved in the variety of notions of "consciousness" now under discussion in various disciplines, and attempting to provide clarification through architecture-based analysis, e.g. by showing how different types of consciousness are supported by different architectures; this includes explaining qualia in terms of causally indexical predicates used by a meta-management system (includes work done with help from Ron Chrisley: e.g. Sloman & Chrisley(2003)).[12]
- Exploring the variety of forms of representation and their trade-offs, and in particular the use of spatial or diagrammatic thinking and reasoning (mainly attempting to analyse in some detail what the requirements for doing that are and how they relate to visual architectures and their functions).[13]

[10] A cluster concept C has two main features: its instances possess some subset of a collection of features, F1, F2, F3, ... associated with C, but there is no determinate definition of C in terms of required features from the set. Thus people using the concept may disagree as to which features are necessary or sufficient, and even an individual may be undecided about this, without having any clear way to decide on indeterminate instances.

[11] See **http://www.cs.bham.ac.uk/research/cogaff/sloman-chrisley-scheutz-emotions.pdf**

[12] See also this invited talk at ASSC7 **http://www.cs.bham.ac.uk/research/cogaff/talks/#talk25**

[13] An incomplete draft paper is available: **http://www.cs.bham.ac.uk/research/cogaff/sloman-diag03.pdf** Previous papers on forms of representation, including (Sloman, 1971). are also on that web site.

- Analysing the concept of *possibility* used in causal thinking and in perception of affordances – e.g. seeing the possibilities for action in a scene, and the constraints on possible actions, and also more abstract affordances involved in the mental states of others – e.g. seeing someone as angry. This has architectural implications.[14]
- Other informal work on ontologies for intelligent agents embedded in a physical world with or without other agents, and ontologies for scientists studying such agents (with Ron Chrisley (Sloman and Chrisley, 2005)).
- Work done with Brian Logan on problem-solving (especially planning) subject to multiple, possibly incommensurable, constraints:[15] this is related to my earlier work on the logic of "Better" (Sloman, 1969). The planning work included problem-solving at different levels of abstraction and combining spatial and non-spatial reasoning.
- Work done with Matthias Scheutz on varieties of arbitration (conflict detection and resolution) in complex multi-functional architectures. The point is that instead of a monolithic arbitration mechanism (as in contention scheduling) all sorts of different kinds of conflict resolution are needed in different parts of the architecture. (Compare Minsky's *Society of Mind*).
- Work done with Matthias Scheutz[16] on exploration of small regions of design space by running simulated evolutionary/Alife experiments in which simple agents with partly similar designs compete – illustrating one method for exploring relations between design space and niche space in small "neighbourhoods" (this uses Scheutz' Simworld package illustrated here **http://www.nd.edu/~airolab/simworld/**).
- Work done with Brian Logan and other colleagues on requirements for distributed agent simulations. The Nottingham team, in collaboration with Theodoropoulos at Birmingham are investigating a distributed version of SimAgent based on HLA.[17]
- Relationships between design space and niche space, and the importance of interacting trajectories in both spaces in both evolution and individual development (Sloman, 1994; Sloman, 2000b)
- Characterising the sub-space of design space relevant to individual intelligent agents with integrated bodies, embedded in a physical (or simulated physical) environment, having to cope with multiple dynamically changing goals and many resource limits: this has led to the CogAff architecture schema described below.
- Trying to identify constraints on architectures for human-like agents, including constraints that arise from requirements provided by the agent's context and goals, constraints that arise from an evolutionary history, and constraints that arise from the requirements for individual development: this has led to the H-Cogaff architecture (for human-like agents) described below, a special case of CogAff.
- The architecture and functionality of a human-like visual system within the context of a multi-layered agent architecture, taking explicit account of many requirements usually ignored in research on vision, including the requirement to perceive and understand

[14] A paper on this was presented at KR96 (Sloman, 1996), and the topic is being developed in various online papers and presentations.

[15] Reported in (Logan and Alechina, 1998; Alechina and Logan, 2001)

[16] E.g. see (Scheutz and Sloman, 2001; Scheutz and Logan, 2001; Scheutz, 2002)

[17] See (Lees et al., 2003a; Lees et al., 2003b; Lees et al., pear)

affordances and the requirement of a visual system to be integrated with a multi-functional central system.[18]

- Attempting to understand the nature of human mathematical capabilities, including attempting to relate such capabilities to the existence of a meta-management architectural layer. This is a development of Chapter 8 of (Sloman, 1978) which discusses how a child learns about numbers. One of the issues to be addressed is what capabilities in the architecture make it possible for a child to grasp various kinds of generalisations, for instance: generalising finite sequences of numbers to the idea of an *infinite* sequence; generalising various kinds of changes in perceived lines to the limiting case of a *Euclidean line* which is infinitely straight, perfectly thin, unbounded, etc.; generalising the notion of area of an array of squares that can be counted to the notion of the area of an arbitrarily shaped region. I conjecture that all of these depend on an aspect of the meta-management architectural layer.
- An architecture-based analysis of the nature of semantics, exposing the errors in the popular notion of 'symbol grounding' whether as the specification of a task or as a form of explanation. In particular for some kinds of meaning complex prior structures are required in order that sensory input be capable of playing any role (as first argued by the philosopher Kant). See **http://www.cs.bham.ac.uk/research/cogaff/talks/#talk14**
- Attempting to understand the trade-offs between *precocial* species (or designs) that produce individuals (like horses, deer, chickens, insects) whose innate (i.e. mainly genetically determined) capabilities enable them to be to a considerable degree self-sufficient from the start, and *altricial* species (or designs) that produce helpless individuals (e.g. newborn or newly hatched lions and other hunting mammals, nest-building birds, monkeys, apes and humans) that require extended help and support from older carers. The conjecture is that the skills and know-how required for the altricial animals in their adult life have a kind of richness, complexity and variety, that may not fit within the resources available to a genome: instead evolution 'discovered' the advantage of specifying architectures and mechanisms that develop after birth through complex interplay between innate (boot-strapping) mechanisms and interactions with the environment. This argument assumes that it would not be feasible to build something like a precocial infant and then merely add "software" content through various processes of learning. However there are unanswered questions about the nature of the boot-strapping process. This related to the attack on symbol grounding, mentioned above.
- Attempting to understand requirements for various forms of learning, including the following :
  - developing or extending forms of representation and ontologies
    (including learning new languages, types of diagrammatic representation, types of mathematical formalisms, programming formalisms, etc. and the sorts of things they can represent)
  - learning to perceive/recognize new classes of objects, events, processes, relationships,
  - acquiring knowledge of individuals – both those that are immediately present and others that endure when not perceived

[18] Some of this work, begun in the 1970s was reported in (Sloman, 1978) and elaborated recently in the online version. A recent presentation on this is here **http://www.cs.bham.ac.uk/research/cogaff/talks/#talk21**

    - acquiring knowledge about spatial locations and relationships of particular objects and places
    - acquiring general knowledge (e.g. causal and statistical laws)
    - acquiring new reasoning abilities
    - acquiring new results of reasoning, planning and problem-solving that can be stored for re-use.
    - acquiring new behavioural skills
    - transferring skills from one part of the system to another: e.g. a deliberative mechanism training a reactive mechanism, or a reactive mechanism's performance producing generalisations that can be learnt by a deliberative mechanism,
    - acquiring new goals, preferences, standards, values, attitudes.
    - learning new strategies for conflict resolution or behaviour arbitration,
    - learning new forms of self observation: including introspective skills and observation of aspects of external behaviours, e.g. effects of one's behaviour on others.
    - learning new social skills including new ways of interpreting actions, expressive behaviours and linguistic utterances of others, and new ways of communicating with or influencing others.

  (That is not a *definitive* or *complete* list, merely a collection of different types of learning that might be expected in a human-like system.)

- Describing weaknesses in the standard functionalist analysis of mind (which we call *Atomic State Functionalism* (AFM)) and showing how those weaknesses can be avoided by adopting a different theory *Virtual Machine Functionalism* (VFM). (For more on this see the ASSC7 presentation: **http://www.cs.bham.ac.uk/research/cogaff/talks/#talk25**)
- Relating the philosopher's notion of "supervenience" to the software engineer's notion of "implementation" including showing how clarification of the notion of "virtual machine" can help both with philosophical and with scientific problems (with help from Matthias Scheutz). In particular, this analysis deals with the problem of how events and processes in virtual machines can have causal powers.[19]
- Analysing requirements for tools for research on these topics and, with help from Brian Logan, Matthias Scheutz, Darryl Davis, Catriona Kennedy and other users. designing, implementing, testing and extending SimAgent, a sophisticated multi-paradigm integration platform to support rapid prototyping and rapid deployment of experimental instances of different sorts of architectures in a variety of scenarios. On the basis of this experience I have submitted (with Brian Logan, Rick Cooper and others) a collaborative grant proposal for a far more ambitious version: **http://www.cs.bham.ac.uk/ axs/basictech**
- Attempting to design a large scale multi-disciplinary project that can stretch our understanding of these issues and also produce useful practical results of many kinds (including improved forms of education, therapy and counselling, based on a theory of the architecture of a learner or patient): the proposal is to focus on a stage in human development at which rich, structured, interactions with other individuals and with the environment are possible (e.g. a child between 2 and 5 years of

[19] Discussed in the IJCAI-01 tutorial presented with Matthias Scheutz **http://www.cs.bham.ac.uk/ axs/ijcai01** and in draft papers and slide presentations **http://www.cs.bham.ac.uk/research/cogaff/talks/#talk12** and **http://www.cs.bham.ac.uk/research/cogaff/talks/#talk22**

age) in order to identify the collection of culture-neutral capabilities on the basis of which a huge variety of types of adults can develop through personal histories and cultural influences. This work has benefited from interaction with Marvin Minsky and Push Singh at MIT who developed related ideas independently, The current draft proposal, being discussed under the auspices of the UK Computing Research Council is here **http://www.cs.bham.ac.uk/research/cogaff/gc/**. An earlier, more detailed proposal developed partly in discussion with Minsky and Singh is here **http://www.cs.bham.ac.uk/research/cogaff/manip/**

- Attempting with Ron Chrisley and Matthias Scheutz to analyse general notions of "information", "information-user", and the related more general and basic notions of "need" and "function". This work is done with the aim of explaining why it is correct to describe all biological orgnanisms, even the simplest, as using information, in the sense of "information" that implies having semantic content. If we can clarify these ideas we can also explain in what sense non-biological machines can use information, countering both arguments (e.g. by Dretske, Millican, and others) claiming that notions of function and of information presuppose some kind of evolutionary origin for the systems involved, and also arguments (e.g. by Dennett, Newell and others) that in order to treat something as an information-user it is necessary to take up an "intentional stance" (attributing rationality to information users) since there is no objective sense in which anything uses information comparable to the objective sense in which something conducts electricity, for instance. Our counter-analysis based on what Dennett called "the design stance" is presented both in various slide presentations here http://www.cs.bham.ac.uk/research/cogaff/talks/ and in a recently completed paper (Sloman, Chrisley, Scheutz, To Appear). One reason why the intentional stance is not required is that we regard very primitive organisms, including for instance bacteria as well as instance, as using information, whereas it is inappropriate to describe them as being either rational or irrational. The same can be said of many artificial control systems, including such simple devices as thermostats.

I have listed this diverse (some would say chaotic) collection of topics because my work (including much collaborative work) has actually been concerned with all of them, but more importantly because I believe that projects that aim to understand and perhaps simulate or replicate animal intelligence, including human intelligence, will need to investigate all of these issues, and probably others not listed.

In contrast most researchers focus on a small subset – quite understandably, given that they generally have to produce project reports in a two or three year time scale, using only small research teams. whereas The work described here is far more long term and requires close collaboration with and help from people in many disciplines.

1.4 The need for integration of diverse capabilities

The list of points above gives an indication of the variety of concerns to be addressed in developing an adequate theory of the sort of information-processing architecture required for a human-like system. We need to understand the variety of human competences and their mutual dependencies. For practical reasons, most AI research can be seen as attempting to take some arbitrary subset of the collection of human abilities and replicate them in machines, e.g. parts of language understanding, parts of visual perception, certain kinds of planning, certain kinds of mathematical reasoning, a subset of types of learning, some emotional behaviours, some forms of motor control, etc. It is not obvious that such selective projects can succeed: it may be the case, and I suspect it is the case, that each of those human capabilities depends essentially on a large core of architectural features, representational abilities, and mechanisms shared by all of the abilities and extended by the abilities. For instance, the kinds of things we learn to see can depend on the kinds of goals we have; and what we learn to see extends the ontology available for formulating both goals and plans; these in turn affect what we are able to think and say using a human language, which can allow us to formulate questions that lead to new goals, new visual abilities, new reasoning powers, etc. All of this is clearly illustrated in the history of science and culture: the conjecture is that something similar happens in the development of individual humans.

If that is correct, attempts to solve the problems in isolation may forever lead only to brittle, incomplete, systems with limited usefulness, either as explanatory models or as applications of AI.

This diagnosis is offered as an alternative to more common diagnoses that blame restricted progress in AI on the use of particular mechanisms or forms of representation (e.g. criticisms of "GOFAI") or the development of disembodied systems. The point that is often missed by such critics is that alternative approaches that also focus on a limited subset of abilities may also fail to make significant progress, except in restricted domains.

2 How to proceed?

Let's assume we are trying to understand how to build a system something like a human being. We are not *primarily* concerned with the task of building one: people make new human beings all the time. Understanding is our main goal. However, this understanding could be used in building, and attempts at building will contribute to our understanding. But understanding involves acquiring new knowledge. That presupposes that there are gaps in our knowledge. In order to make progress we have to understand the nature of those gaps. This will help us understand the questions we should be asking.

2.1 Which questions should we try to answer?

What sorts of questions do we need to ask in order to make progress? There are many sorts of questions that are not helpful, because they make false assumptions or ignore conceptual confusions and ambiguities. Examples of such misleading questions are these, if asked about a whole human or human-like robot:

- Which algorithm does it run?
- Which formalism does it use?
- What learning mechanism does it use?
- What is the knowledge required in a human-like system?
- What are the capabilities of a human-like system?
- What sort of architecture does it need?
- What sorts of physical mechanisms are adequate for the implementation?
- Does it need to have emotions?

- What is attention?
- How much computing power does a human-like system need?
- What sort of body does it need?

We can juxtapose those questions with more productive questions:

- Instead of: *Which algorithm does it run?*
  Ask: which sorts of algorithms are sometimes useful, and in which portions of the architecture? What are they good for, and what are the trade-offs between the alternatives?
- Instead of *Which formalism does it use?*
  Ask: what sorts of formalisms are possible, and in which portions of the architecture can they be used? What are they good for, how do they differ, and what are the trade-offs between the alternatives?
- Instead of *What learning mechanism does it use?*
  Ask: how many different varieties of learning are possible, what are they useful for and which portions of the architecture can or should use them?
- Instead of *What is the knowledge required in a human-like system?*
  Ask: what are the different sorts of knowledge that different sub-systems in an intelligent system may need, what are they needed for, and what are the trade-offs between the alternatives? Is the same knowledge needed in different forms in different sub-systems?
- Instead of *What are the capabilities of a human-like system?*
  Ask: What are the different sorts of capabilities that human like systems may need at various stages in their development, or in different contexts, or in different subsets of their architectures, and what roles do those capabilities play? Which capabilities need to be present for the others to develop, or be controlled, or be modified as needed to suit new contexts?
- Instead of *What sort of architecture does it need?*
  Ask: What varieties of architectures can be more or less human-like, and which sorts are typically present in humans at different stages of development? What are the features of different architectures that make them appropriate to different physical environments, different cultures, different stages of development, or different tasks and sub-tasks in those various stages and contexts?

  For architectures which have large scale sub-structures with their own architectures, e.g. perceptual sub-architectures, motor sub-architectures, memory sub-architectures, it can be useful to repeat all the above questions for each sub-architecture.
- Instead of *What sorts of physical mechanisms are adequate for the implementation?*
  Ask: What sorts of virtual machines are required to support the various components and capabilities, and how are those virtual machines implemented in the underlying physical mechanisms? E.g. how many layers of virtual machines are required?
  (It is possible that new advances in physics or in brain science will draw our attention to types of information processing mechanisms as yet undreamt of.)

  For the various kinds of virtual machines that seem to be potentially useful, which sorts of physical mechanisms are sufficient to provide implementations? Are there some mechanisms developed by evolution that cannot be replicated using currently known processes for building computers?
- Instead of *Does it need to have emotions?*
  Ask: How many varieties of affective states, processes and mechanisms are possible? Which of them are needed in various portions of a complete architecture? How many of them are not based on distinct mechanisms but arise out of interactions between other mechanisms?

(It is not obvious that we have a good ontology for talking about possible collections of affective components. Whether various kinds of emotions will be there as a set of required components or merely as 'emergent' manifestations of the operation of other components, remains to be seen. In part it will depend on how 'emotion' is defined.

- Instead of *What is attention?*
  Ask: What sorts of attention functions are required in the different subsystems?
  For instance, visual attention switching is partly done by changing the direction of the whole body, or by changing the direction of the eyes (foveating), or by a *purely internal* switch of processing to a subset of the visual field, or by switching the type of processing done (e.g. attending to an object's shape, *vs* its colour, *vs* its texture, *vs* the material of which it is made *vs* its fragility or stability or graspability or reachability, etc.) Some of these switches are deliberate and voluntary. Some are conscious but involuntary, for instance when attention is drawn to a movement. Probably there is a vast amount of switching of resources, selection of information, selection of type of processing that goes on unconsciously in the reactive sub-mechanisms (e.g. concerned with actions like grasping, walking, riding a bicycle, driving a car, sight-reading music, etc.). The different forms of attention may use very different mechanisms. E.g. there are differences between attention control by physical movement, by inhibition of competing sub-processes, by "switching on" a normally inactive disposition, by modifying the mode of operation of an ongoing process.
- Instead of *How much computing power does a human-like system need?*
  Consider this: When we know more about what brains do, this question may begin to make sense, but at present it is unanswerable. If much of the computing power of brains is at the level of chemical information processing (molecular interactions) then estimates based on counts of neurones could be wrong by several orders of magnitude. A recently discovered fact that may have great relevance to these questions is the fact that biological neural nets can change their architecture (i.e. their connectivity) in milliseconds.
- Instead of *What sort of body does it need?*
  Ask: in how many different ways can a human-like system be embodied: what are the consequences of different sorts of embodiment, e.g. for constraints on what the system can do and for providing new opportunities for learning, thinking, doing, etc.? What difference does it make whether a system is implemented with a physical body or merely a simulated physical body in a simulated physical environment? If a physical implementation abstracts from details of biological bodies what differences will it make to the mental functioning of the system?
  (Consider humans born with various physical limitations or defects, e.g. the limbless "thalidomide babies" of the 1960s, or people who suffer injury and are provided with non-biological prostheses.)

Summary: if we ask the wrong questions because of over-simplified presuppositions, we'll not find the important answers.

2.2 Resources and resource limits

One kind of question often addressed in designs for computational systems is: How can resource limitations be overcome or their impact minimised? In the world of software the two most commonly addressed resources are space (or, to be more precise, available physical or virtual memory) and time. Thus many design decisions are concerned with reducing either memory requirements or time requirements for accomplishing a task, or both – though often there is a trade-off between space and time. Another resource often referred to in discussions of organisms is energy: brains consume energy and the amount of energy consumed at any time depends on what the brain is doing. Yet another may be the neurotransmitters and other chemicals in the brain that are consumed more rapidly during intense activity, and may in some circumstances be consumed faster than they can be produced. This may be why so-called "arousal" states cannot be left on permanently even though they improve performance of various kinds.

But there are other resource limitations that are not normally noticed. The H-Cogaff architecture, sketched below, can be seen as a solution to several different kinds of resource limitations.[20]

Besides computational resource limits which have impacts on designs, there are other resource limits (mentioned only briefly in my papers and the 1978 book) including things like:

- having only (at most) two hands,
- being in only one location,
- not being able to look in all directions at once,
- various other physical limitations of an organism or agent: strength, speed, food storage capacity, etc.
- Limitations of various kinds of resources in the current environment, or resources in the accessible parts of the environment, including:
  - food,
  - construction materials,
  - shelter,
  - mates,
  - information.

Just as physical limitations may be associated with physical impossibilities, such as being in two places at once, or moving in two directions at once, so may information processing limitations arise from incompatibilities between types of internal processes, for instance not being able concurrently to do certain kinds of reorganisation of information while functioning in a demanding environment where that information has to be used. Familiar examples include not using a file-store while it is being backed up, not running any processes during a garbage collection, and not running procedures while they are being recompiled. (Perhaps this has something to do with the need to sleep in some animals.)

2.3 Resource limits connected to evolutionary history

One kind of resource limit that has driven aspects of the design of the H-Cogaff architecture arises from the notion that evolutionary history is a kind of resource that provides useful information.

Some organisms have a kind of niche for which their evolutionary history provides almost all the necessary information, for instance in the form of genetically compiled behaviour rules which combine with sensory inputs to produce behaviours – possibly alongside simple forms of adaptation where appropriate. However, some organisms may not have evolutionary histories that provide enough information of this kind.

(a) For instance, there are organisms, especially those that use their intelligence to change the environment, whose niches evolve in complex ways for which evolutionarily selected behaviour rules are not adequate: evolution takes too long to catch up. Or there may not have

[20] This section arose out of recent email discussions with Andrew Coward, following a brief meeting at ASSC7

been opportunities in the past during which appropriate behaviours could have evolved, e.g. behaviours for driving cars and trucks, or flying aeroplanes.

(b) Another possibility is that the organism's niche involves such a great diversity of of possible goals, problems, plans and solutions that it would not be possible for evolution to have time (even over several billion years) to discover all the problems, discover solutions to them and encode them in genes.

(c) A further possibility is that the variety of possible cases is too large to be encoded in the genome even if they could somehow all be explored in advance, given that the molecules involved in biological reproduction have size limits. For instance, the egg cells and sperm cells and the information structures that specify what should go into them must be small enough to fit in the parents while they are being built.

In all of those cases the genes will not encode a wide enough collection of behaviours. To deal with this, evolution appears to have 'discovered' a powerful alternative to evolution: and that's how deliberative systems came about. For they are systems with the ability to create *novel* solutions when faced with new problems: though they will do so by combining or modifying old solutions or components of old solutions.

This has two distinct aspects, depending on whether searches for suitable behaviours are made in the environment or in the mind of the organism using processes of simulation or more abstract representation.

- *Explorations in the environment:*
  The first case requires the ability to try out various combinations of actions in the environment in order to find which ones will work. Doing that systematically requires information-processing resources that include the ability to remember the results of different attempts, so that a recent attempt and an older attempt can be compared and one of them chosen as better. This requires an appropriate ontology for describing complex combinations of actions and for using the description to generate an action sequence.

  In simpler cases the environment can provide the memory: for instance, if you try to build two kinds of climbing aid for getting to food in inaccessible places, or two different kinds of cutting tools, you can keep both devices, compare them and then use one (or perhaps use different ones in different contexts). In other cases one attempt involves undoing the results of another attempt: e.g. if parts have to be re-used. This increases demands on memory within the organism. Later this could be supplemented by external symbolic records of the different attempts.

- *Explorations in the mind:*
  In some cases it is too risky, or too time consuming to try out options by acting in the environment. For instance failure could mean death. In such cases it is useful to be able to manipulate *representations* of the different possible sequence of actions and then compare the results of those manipulations in order to select the best combination. This is sometimes referred to as simulating or modelling actions instead of performing them. In certain cases it may be possible to do this using external representations (e.g. drawing diagrams in the sand, or miming certain actions to see what their consequences would be). However in some cases it can be done entirely internally: e.g. visualising a route, or a way of manipulating some object.[21]. This is what I have been referring to as a fully fledged *deliberative* capability, although various subsets of such a deliberative system are also possible, and are probably to be found in many non-human species and in very young children.

[21] The benefits of experimenting with a mental model or simulation instead of with reality have been pointed out by many people: including Craik, Popper, Dennett, ...

We don't yet understand the full implications of all of these alternative modes of exploring and comparing possibilities in order to select an action. The information-processing requirements are different depending on which of the various means are used. Moreover we don't know what mechanisms can be used in animal brains to implement these techniques nor what the costs are. In particular, in biological systems the cost in brain mechanisms and energy and other requirements for a fully deliberative system appear to be very high: that's why there are relatively few organisms that have such things: they have to be atop a food pyramid.

Another resource limit is the storage capacity of genes. Evolution discovered that by altering processes of growth and development in individuals and providing bootstrapping mechanisms to extract information from the environment, it could reduce the requirement to store precise details (including parameters that match the physical variations in individuals produced from the same gene pool).

This explains the difference between altricial and precocial species. Humans are the extreme case of the former but there is much we don't understand (e.g. how the bootstrapping process actually helps to create the architecture to fit the environment).

[Add some more on subtle resource limits: e.g. limited opportunities for individuals to learn may be compensated for by cultural transmission. Work in the social sciences is relevant here.]

3 The importance of architectures

AI has always been concerned with algorithms and representations, but we also need to understand how to put various parts together into complete working systems, within an *architecture*. It is now common in AI and Cognitive Science to think of humans and other animals, and also many intelligent robots and software agents, as having a virtual machine information processing architecture which includes different layers, and which, in the case of animals, evolved at different stages. But many different architectures are proposed, and there is no clear framework for comparing and evaluating them.

Part of the framework should be a specification of what architectures are for: what requirements they have to meet. Each set of requirements can be thought of as a *niche* in which different architectures may fit more or less well. The set of possible niches for organisms or machines is therefore the set of possible sets of sets of requirements, referred to in our papers as "niche space" since Sloman(1994).[22] The previous section has indicated some of the features of niche space for biological organisms. We also need to understand the space of possible designs, *design space*, as well as the space of possible relationships between designs and niches (for which the widely used notion of a "fitness function" is not generally adequate).

Exploration of a space of designs requires the use of an ontology for designs. This will include many components of designs, including hardware mechanisms, algorithms, forms of representations and architectures within which different functional components are integrated. One of the requirements for exploring design space is therefore having a language for talking about architectures. Unfortunately there is not yet any agreed ontology or language: individual researchers invent labels and diagrams to present their ideas but there are many ambiguities and

[22] I recently discovered that some biologists use the phrase "niche space" in a different way: they talk about the niche space of a particular type of organism, namely the range of variation within which that type of organism can cope. For examples of this usage see http://www.csuchico.edu/˜pmaslin/limno/niche.space.html and http://research.esd.ornl.gov/˜hnw/esri99/

confusions.

Often two architectures (e.g. two multi-layer architectures) may appear to be very similar at first sight, e.g. because similar diagrams are used, though detailed analysis shows important differences. For example, many researchers talk about reactive, deliberative and reflective components of architectures, but in fact are using those labels to have quite different meanings. Conversely presentations that appear to be concerned with different sorts of architectures may actually be representing the same thing, but with different emphases, different notations, etc.

3.1 Deceptively similar labels

Some use "reactive" to mean "stateless". Some use it to refer to innate architectures that can merely learn using reinforcement mechanisms. Some allow "reactive" to include quite complex symbolic processes[23]

Likewise the word "deliberative" turns out to have different connotations for different authors, and that is partly because the space of possible architectures is so rich and we still lack comprehensive surveys, so people focus on different boundaries within it. For instance some people will label as "deliberative" anything that can execute symbolic instructions, or anything that is capable of being presented with a choice and selecting an option, as in (Arbib, 2002) or in contention scheduling architectures (Cooper and Shallice, 2000), discussed further below. For a long time I naively thought everyone agreed that "deliberative" referred to a complex mixture of capabilities including

- The use of structured representations with compositional semantics
- The ability to represent and compare and evaluate non-existent but possible entities, or alternative answers to questions about what existed in the past or what exists at some invisible location
- The ability then use the selected alternative to perform some task with it, e.g. execute a plan, predict some future events, explain an observed phenomenon, etc.
- The ability to learn generalisations about the environment and store them for future use in order to extend those deliberative capabilities,
- In particular, the ability to observe the effects of executing plans so as to learn out to do things differently in future (which I think is what Minsky refers to as 'reflection').
- And possibly to store the results of particular "deliberations" for future use, e.g. as reusable plans, perhaps in an abstracted form (as in Sussman's HACKER (Sussman, 1975) and in SOAR).

These capabilities in their fullest form have very demanding requirements, especially if implemented in animal brains, and that, perhaps is why the full set of such capabilities evolved very late and appears to be very rare among animals, although simple versions of such systems proved relatively easy to implement on computers, as happened in early AI systems.

However, mainly as a result of working with Matthias Scheutz (Scheutz and Sloman, 2001; Scheutz and Logan, 2001; Scheutz, 2002) on a variety of intermediate systems which go beyond what I mean by "reactive", i.e. with some ability to represent and compare unobserved hypothetical situations or actions, but do not possess the full set of deliberative capabilities listed

[23] E.g. See Nilsson et al. (1994) on teleoreactive systems , and their online demonstration of a working teleoreactive system:
http://www.robotics.stanford.edu/users/nilsson/trweb/TRTower/TRTower_JAVASCRIPT.html

above I have come to realise that we must replace our existing terminology in order to allow for a much richer variety of possible intermediate mechanisms between the simplest reactive systems and the full set of human-like deliberative capabilities.

This will also be relevant to guiding the search among biologists for different sorts of explanations of animal behaviour and also the search for theories about evolutionary trajectories, and possibly also developmental trajectories in individuals, if architectures are not formed at birth but develop during interactions with the environment (in altricial species).

Yet another ambiguity concerns the label 'reflective', often used as a label in descriptions of multi-layered architectures. For instance many people refer to a type of three-layered system as containing a reactive layer, a deliberative layer and a reflective layer; and in addition to previously mentioned ambiguities in the first two labels there are ambiguities in the third. For som researchers, 'reflective refers to the ability of an information-processing system to observe and perhaps control some of its own internal states and processes (as in discussions of reflective programming languages). For others it refers to any ability of a system to observer its own behaviours whether internal or external. Thus in that sense any system with feedback control mechanisms would be reflective. Alternatively it may be restricted to a system's observations of its behaviour that are not immediately used to control behaviour, as happens in very many feedback systems, but instead are used to form generalisations or detect faults that can lead to long term changes in plans or skills.

About ten years ago Luc Beaudoin suggested the use of the label 'meta-management' for the processes in a system that could observe, learn from and modify some of its own processes of deliberation and goal selection, which we later generalised to include observation of other internal phenomena, for instance intermediate structures in perceptual processing, which I have used to provide an architecture-based analysis of the philosophers' notion of 'qualia' (most recently elaborated in (Sloman and Chrisley, 2003)). For several years I thought that our label 'meta-management' and the label 'reflective' used by many others were more or less synonymous. However it is now become clear that that is not so: there really are different uses. This also explains why Minsky distinguishes 'reflective' (which labels some of the phenomena that I have included under 'deliberative' above) and 'self-reflective' (which refers to a subset of the phenomena that I have been calling 'meta-management').

There are other ambiguities regarding what people mean by perception, or more specifically vision, for instance where they draw boundaries between perception and cognition.

All this is in addition to the more glaring and obvious confusion and ambiguities involved in words like 'emotion', 'affect', and 'consciousness'.

This terminological confusion may or may not be seriously holding up communication, criticism, collaboration and progress among researchers. I suspect that when people think they agree because they use similar labels that can seriously hamper progress.

3.2 Ambiguity of "architecture"

Some computer scientists still use the word 'architecture' only to refer to the physical or digital electronic architecture of a computer, as was common about 20 or 30 years ago, and still is in courses on computer architectures. However the word can also be used to refer to the architecture of a company, a symphony, a compiler, operating system, a theory or a mind. In particular, it can be used to describe any complex system made of coexisting parts which interact causally in order to serve some complex function or produce some behaviour.

The parts may themselves have complex architectures. The system and its parts need not be physical. Nowadays the word often refers to non-physical aspects of computing systems, i.e. *virtual machines*. E.g. an operating system or chess program is a virtual machine with an architecture, though it will need to be implemented in a physical system, usually with a very different architecture.

3.3 Ambiguities of 'Information processing'

'Information processing' is another term which has both narrow and broad interpretations: some people restrict it to refer to the kinds of bit-manipulations that computers do. However it can be used to refer to a wide range of phenomena in both discrete and continuous virtual machines of various kinds, including acquiring perceptual information about an environment, storing facts, deriving new consequences, searching a memory or database for answers to questions, creating plans or strategies, generating goals, taking decisions, giving instructions or exercising control. As the last two illustrate, not all information is *factual*: there is also *control* information, including very simple on-off control signals, variations in continuous control parameters, labels for actions to perform, and descriptions of what is to be done.

Explicit or implicit theories of mental architecture are not new. Early empiricist philosophers thought of the mind as a collection of 'ideas' floating around in a sort of spiritual soup and forming attachments to one another. Kant (Kant, 1781) proposed a richer architecture with powerful innate mechanisms that enable experiences and learning to get off the ground, along with mathematical reasoning and other capabilities. Freud's theories directed attention to a large subconscious component in the architecture. Later Craik proposed (in 1943) that animals build 'models' of reality in order to explore possible actions safely without actually performing them. Popper (in (Popper, 1976) and earlier works) advocated similar mechanisms which, he said, would allow our mistaken hypotheses to 'die' instead of us. Recent work has added more detail. Albus (p.184 of (Albus, 1981)) depicts MacLean's idea of a 'triune' brain with three layers: one reptilian with one old and one new mammalian layer. A neuropsychiatrist, Barkley, has recently begun to develop a sophisticated architectural model, partly inspired by J. Bronowski, to account for similarities and differences between normal human capabilities and sufferers from attention disorders (Barkley, 1997), though most psychologists and neuroscientists find it very difficult to think about virtual machine architectures. Shallice and Cooper are among the exceptions (Cooper and Shallice, 2000).

In the meantime, AI researchers have been exploring many sorts of architectures. See Nilsson's account ((Nilsson, 1998), Ch 25) of *triple tower* and *triple layer* models. Architectures like SOAR, ACT-R, and Minsky's *Society of Mind* have inspired many researchers.

But too many researchers choose an architecture, produce some arguments about its 'advantages', produce an implementation and then give some more or less impressive (or unimpressive) demonstrations of what it can do.

That is not an adequate way to do science. It is as if physicists or chemists were to study a few physical elements and a few chemical compounds while ignoring all the rest.

We can compare that with a field that studies architectures, but has no general overview of the space of interesting or important architectures, or the different types of requirements against which they can be evaluated, though Dennett (Dennett, 1996) makes a good start. In short, there are no adequate surveys of 'design space' and 'niche space' and their relationships (described briefly in (Sloman, 2000b)).

As a first-draft partial remedy, we offer the CogAff schema depicted in figures 1(a), (b) and 2(a), (b), and described below. This is not an architecture, but a generic *schema*, something like a *grammar* which covers a variety of different cases. Just as a grammar distinguishes types of sentence-components and specifies how they can be combined, so our architecture schema, distinguishes types of components of the architecture, e.g. reactive, deliberative and meta-management, and ways in which they can be combined. On that basis we can begin to produce theories of the kinds of states and processes that can be supported by the various architectures, for instance and showing how the richer instantiations of CogAff are capable of having at least three different classes of emotions and much besides.

Figure 1: (a) (b)

The CogAff architecture schema combines cognitive and affective components. Nilsson's 'triple tower' model, with information flowing (mainly) through perceptual, central, and motor towers, is superimposed on his 'triple layer' model, where different layers, performing tasks of varying abstractness, use different mechanisms and representations. In (a) the two views are superimposed. Components in different boxes have functions defined by their relationships with other parts of the system, including information flow-paths (not shown). In (b) a fast (but possibly stupid) alarm system receives inputs from many components and can send control signals to many components. An insect's architecture might include only the bottom layer. Some animals may have reactive and deliberative layers. Subsumption architectures have several levels, all in the reactive layer. Humans seem to have all three layers. See the text and fig. 2 for details. The diagrams leave out some components (e.g. motive generators) and some information pathways, e.g. 'diagonal' routes.

4 Information processing architectures

It is also now commonplace to construe many biological processes, including biological evolution and development of embryos as involving acquisition and use of information. Perhaps the biosphere is best construed as an information processing virtual machine driven partly by co-evolutionary interactions.

5 Conceptual confusions

A problem surrounding the study of architectures is the diversity of high level aims of AI researchers. Some try to solve engineering problems and care only about how well their solutions work, not whether they model natural systems. Other researchers attempt to understand and model humans, or other animals. A few are attempting to focus only on general principles equally applicable to natural and artificial systems. An effect of all this is that there is much confusion surrounding the description of what instances of the proposed architectures are supposed to be able to do. For instance, someone who describes a system as 'learning' may merely mean that it adaptively solves an engineering problem. Another may be attempting to model human learning, perhaps without being aware of the huge variety of types of learning. An engineer may describe a program as using 'vision' simply because it makes use of TV cameras to obtain information, which is analysed in a highly specialised way to solve some practical problem, ignoring the fact that animal vision has many other aspects, for instance detecting 'affordances' (Gibson, 1979; Sloman, 1989; Sloman, 2001b).

Study of emotion has recently become very fashionable in psychology and AI, often ignoring the vast amount of conceptual confusion surrounding the term 'emotion', so that it is not clear what people mean when they say that their systems have emotions, or model emotions, or use affective states (Sloman and Logan, 2000; Sloman, 2001c; Scheutz, 2002) Social scientists tend to define 'emotion' so as to focus on social phenomena, such as embarrassment, attachment, guilt or pride, whereas a brain scientist might define it to refer to brain processes and widespread animal behaviours. The word has dozens of definitions in the psychological and philosophical literature, because different authors attend to different subsets of emotional phenomena.

McDermott's critique (McDermott, 1981) of AI researchers who use mentalistic labels on the basis of shallow analogies has been forgotten. We offer the CogAff schema as a first-draft framework for describing and comparing architectures and the kinds of states and processes they support. We can then see how definitions of mental phenomena often focus on special cases all of which the schema can accommodate, e.g. as we have shown elsewhere in the case of emotions and vision (Sloman, 2001a; Sloman, 1989; Sloman, 2001b).

5.1 Architecture-based concepts

Understanding the variety of information processing architectures helps to clarify confused concepts, because different architectures support different sets of capabilities, states and processes, and these different clusters characterise different concepts. For instance, the fullest instantiations of the CogAff schema account for at least three classes of emotions: primary, secondary and tertiary emotions, extending previous classifications. (Damasio, 1994; Picard, 1997; Sloman, 2000a; Sloman and Logan, 2000). An architecture-based analysis can lead to further refinements in the classification of affective states. (Sloman, 2001a). Likewise, different concepts of 'seeing' relate to visual pathways through different sub-systems in a larger architecture. 'Blindsight' (Weiskrantz, 1997) could arise from damage to connections between meta-management and intermediate high level perceptual buffers, destroying self-awareness of visual processing, while lower level pathways remain intact.

Architectures differ not only between species, but also while an individual develops, and after various kinds of brain damage or disease. The resulting diversity requires even more conceptual differentiation. ‘What it is like to be a bat’ (Nagel, 1981) may be no more obscure to us than ‘What it is like to be a baby’, or an Altzheimer’s sufferer.

5.2 Cluster concepts

Many of our mental concepts are ‘cluster concepts’: they refer to ill-defined subsets of a cluster of properties. E.g. if an architecture supports capabilities of types $C1, \dots Ck$, then boolean combinations of those capabilities can define a wide variety of concepts. Our pre-theoretical cluster concepts lack that kind of precision; so, for a given mental concept M, there may be some combinations of Cs that definitely imply presence of M, and others which definitely imply absence of M, without any well-defined boundary between instances and non-instances. Cluster concepts may have clear cases at extremes and total indeterminacy in a wide range of intermediate cases, because there has never been any need, nor any basis, for labelling those cases. Worse, we may be unaware of the full range of capabilities (Ci) relevant to clarifying the concept.

When we have a clear view of the space of architectures we can consider the families of capabilities supported by each type of architecture, and define new more precise concepts, just as we have defined primary, secondary and tertiary emotions in terms of reactive, deliberative and meta-management mechanisms e.g. (Sloman, 2001a). Asking which definitions are *correct* is pointless, like asking whether mathematicians are ‘correct’ in defining ‘elliptical’ to apply to circles. Wheel-makers need a different concept.

Some architectures may support all the mental concepts we normally apply to humans. Others may support only simplified forms e.g. ‘sensing’, but not all of our notions of ‘pain’, ‘emotion’, ‘consciousness’, etc. An insect has some sort of awareness of its environment even if it is not aware that it is aware, because there is no meta-management.

If we had a clear idea of the information processing architecture of a foetus at different stages of development, then for each stage we could specify concepts that are relevant. New-born infants, like insects, are limited by their architecture: e.g. they may be incapable of puzzlement about infinite sets or the mind-body problem. Likewise, when describing AI systems, we need to be careful not to over-describe simplified architectures.

If we have a well-defined space of possible architectures, and can investigate precisely which concepts are applicable to which subsets, we can develop agreed terminology for describing agents.

6 What sorts of architectures?

We cannot (yet) hope for a complete survey of possible information processing architectures since we are so ignorant about many cases, e.g. animal visual systems. Perhaps evolution, like human designers, has implicitly relied on modularity and re-usability in order to achieve a robust and effective collection of biological information processing architectures. Figure 1 depicts a biologically-inspired framework covering a variety of architectures, with different subsets of components. It makes a three-fold division between perception, central processing, and action, and contrasts three levels of processing, which probably evolved at different times. (More fine-grained divisions are also possible.) Slow central mechanisms and fast environments may generate a need for fast (but possibly stupid) relatively global ‘alarm’ mechanisms. The need

Figure 2: (a) (b)

The 'Human-like' sub-schema H-Cogaff: (a) lists some components supporting motive processing and 'what if' reasoning in deliberative and meta-management layers. Humans seem to have all those mechanisms, perhaps organised as in (b). The alarm sub-systems might include the brain's limbic system. An interrupt filter partly protects resource-limited deliberative and reflective processes from excessive diversion and redirection, using a dynamically varying penetration threshold, dependent on the urgency and importance of current tasks – soldiers in battle and footballers don't notice some injuries. Different 'personae' can control processing at different times, e.g. when at home with family, driving a car, interacting with subordinates, in the pub with friends, etc. Such an architecture has various kinds of information stores, and diverse information routes through the system, only a subset of which are shown.

for speed in detecting urgent opportunities and dangers rules out use of elaborate inferencing mechanisms in an alarm mechanism, though they may exist in a deliberative layer. Alarm mechanisms are therefore likely to be pattern-based, and to make 'mistakes' at times, though they may be trainable.

Architectures may include different subsets of the CogAff schema. Fig. 2 depicts a conjectured human-like schema H-CogAff,[24] but CogAff allows much simpler instances. Insects probably have only the bottom (reactive) layer (possibly with alarms), and much early AI work was concerned only with the middle portion of the middle (deliberative) layer. HACKER (Sussman, 1975) combined portions of the top two layers. SOAR's 'impasse detection' is a type of meta-management. Brooks subsumption architectures (e.g. in (Brooks, 1991)) include multiple control levels all within the reactive layer, and nothing in the other layers. Moreover, architectures with similar components can differ in their communication pathways.

[24] Our terminology is provisional. We refer to CogAff as a *schema* rather than an *architecture* because not every component specified in it must be present in every architecture to which it is relevant: e.g. it is intended to cover purely reactive agents and software agents which merely contain deliberative and meta-management layers. H-CogAff is schematic in a different sense: it is a conjectured architecture for human-like minds where many components are incomplete or under-specified.

6.1 Layered architectures

The idea of hierarchic control is very old both in connection with analog feedback control and more recently in AI systems. There are many proposals for architectures with two, three or more layers, including those described by Albus and Nilsson mentioned previously, subsumption architectures (Brooks, 1991), the ideas in Johnson-Laird's discussion (1993) of consciousness as depending on a high level 'operating system', and Minsky's notion of A, B and C brains.

On closer inspection, the layering means different things to different researchers. Such ambiguities may be reduced if people proposing architectures agree on a broad conceptual framework specifying a class of architectures and terminology for describing and comparing them, as illustrated in the next section.

6.2 Dimensions of architectural variation

We present some dimensions in which architectures can be compared, originally in (Sloman, 2000c).

6.2.1 Pipelined *vs* concurrently active layers

Often (Nilsson, 1998) the layers have a sequential processing function: sensory information comes in via low level sensors ('bottom left'), gets abstracted as it goes up through higher central layers, until action options are proposed near the top, where some decision is taken (by 'the will'!), and control information flows down through the layers and out to the motors ('bottom right'). We call this an 'Omega' architecture because the pattern of information flow is shaped like an Ω. Many models in AI and psychology have this style e.g. (Albus, 1981). The 'contention scheduling' model (Cooper and Shallice, 2000) is a variant in which the upward information flow activates a collection of competing units where winners are selected by a high level mechanism. The CogAff schema accommodates such pipelines, but also permits alternatives where the different layers are all concurrently active, and various kinds of information constantly flow within and between them in both directions, as in fig. 2(b).

6.2.2 Dominance *vs* functional differentiation

In some designs, higher levels completely *dominate* lower levels, as in a rigid subsumption architecture, where higher levels can turn lower level behaviour on or off, or modulate it. Such hierarchical control is familiar in engineering, and CogAff allows, but does not require, it. In the H-CogAff 'human-like' sub-schema (fig. 2), higher levels partially control lower levels but sometimes lose control, e.g. to reactive alarm mechanisms or because other influences divert attention, such as sensory input with high salience (loud noises, bright flashes) or newly generated motives with high 'insistence' (e.g. hunger, sitting on a hard chair, etc.). In animals *most* lower level reactive mechanisms cannot be directly controlled by deliberative and meta-management mechanisms though indirect control through training is possible.

6.2.3 Direct control *vs* trainability

Even if higher levels cannot directly control lower levels, they may be capable of re-training them, as happens in the case of many human skills. Repeated performance of certain sequences of actions carefully controlled by the deliberative layer may cause an adaptive reactive layer to develop new chained behaviour sequences, which can later be performed without supervision from higher layers. Fluent readers, expert car drivers, skilled athletes, musical sight-readers, all make use of this.

6.2.4 Processing mechanisms *vs* processing functions

Some instances of CogAff may use the same kinds of processing mechanisms (e.g. neural nets) in different layers which perform different functions, concerned with different levels of abstraction. Alternatively, diverse functions may be implemented in diverse mechanisms, e.g. neural nets, chemical controls, symbolic reactive rulesystems, and sophisticated deliberative mechanisms with ‘what if’ reasoning capabilities, using formalisms with compositional semantics. The latter might be used to represent remote or hidden entities, past events, and possible future actions and consequences of actions. If those can be categorised, evaluated, and selected this would support planning, finding explanations of past events, mathematical reasoning, and general counterfactual reasoning. Such deliberative mechanisms require temporary workspace containing changing structures – not needed for most reactive systems.

Deliberative mechanisms that, unlike reactive mechanisms, explicitly represent alternative actions prior to selection, might be *implemented* in reactive mechanisms, which in turn are implemented in various kinds of lower level mechanisms, including chemical, neural and symbolic information processing engines, and it is possible that the reliance on these is different at different levels in the architecture. Some kinds of high level global control may use chemical mechanisms (e.g. hormones) which would be unsuitable for intricate problem solving. If it ever turns out that animal brains require quantum computational mechanisms, e.g. for speed, then these mechanisms could also be accommodated within the CogAff framework.

6.2.5 Varieties of representation

Distinctions between different sorts of representations, e.g. logical, qualitative, diagrammatic, procedural, neural etc. are all relevant, since different components of a complex architecture may have different requirements.

6.2.6 Varieties of learning

There is much research in AI and psychology on learning and individual development. CogAff is compatible with many kinds of learning mechanisms in different parts of the system, including neural nets, trainable reactive systems, extendable knowledge stores, changeable motive generators and motive comparators (see below), extendable forms of representation and ontologies, etc. More subtle types of learning and development can include forming new connections between parts of the architecture, e.g. linking new visual patterns either to reactive behaviours as in athletic training, or to abstract concepts, as in learning to read a foreign language or detect a style of painting.[25]

[25] H-CogAff with its many components and many links also makes possible multiple forms of damage and degradation including changes within components and changes to connections.

In humans the meta-management layer is not a fixed system: not only does it develop from very limited capabilities in infancy, but even in a normal adult it is as if there are different personalities 'in charge' at different times and in different contexts. Learning can extend the variety.

6.2.7 Springs of action, and arbitration mechanisms

Architectures can support 'intrinsic' and 'derivative' motives, where the latter are sub-goals of intrinsic or other derivative motives. Architectures differ in the varieties of motives they can generate and act on and how they are generated, and whether they are represented explicitly or only implicitly in control states. They can also differ in how conflicts are detected and resolved. To illustrate, we mention several contrasts.

Some architectures generate all motives in one mechanism receiving information from other components (e.g. near the 'top' of an Omega architecture) whereas other architectures support distributed motive generation, including reactive and deliberative triggering (fig. 2(b)). In some of the latter, motives generated in different places cannot be acted on unless processed by some central system, whereas others (e.g. H-CogAff) allow distributed concurrent motive activation and behaviour activation. In some reactive systems all reactively generated goals are processed only in the reactive layer, whereas in others a subset of reactive goals can be transmitted to a deliberative layer for evaluation, adoption or rejection, and possibly planning and execution.

Architectures also differ regarding the locus and mechanisms of conflict resolution and motive integration. In centralised decision-making all conflicts are detected and resolved in one sub-mechanism, whereas in others, some conflicts might be detected and resolved in the reactive layer, some might be detected and resolved using symbolic reasoning in the deliberative or meta-management layer, and some might be resolved using highly trained motor sub-systems. Deciding whether to help granny or go to a concert, deciding whether to finish an unfinished sentence or to stop and breathe, deciding whether to use placatory or abusive vocabulary when angry, might all be handled by different parts of the system. In some architectures loci of integration never vary, while others change through learning.

Some systems use 'numerical' conflict resolution, e.g. voting mechanisms, while others use rule-based or problem-solving decision systems capable of creative compromises, and some are hybrid mixtures.

6.2.8 'Peephole' *vs* 'multi-window' perception

Perceptual architectures vary. A 'peephole' model uses a fixed entry locus (using simple transducers or more complex sensory analysers) into the central mechanisms, after which information may or may not be passed up a processing hierarchy, as in the Omega model. In a 'multi-window' model (Sloman, 1989; Sloman and Logan, 2000) perceptual processing is itself layered, concurrently producing different kinds of perceptual information to feed directly into different central layers, e.g. delivering more abstract and more large scale percepts for higher layers, while fine control of movement uses precise and continuously varying input fed into the reactive system or directly to motor sub-systems (fig. 2(b)). Perceptual systems also vary according to whether they are purely data-driven or partly knowledge-based, and whether they can be affected by current goals. Empirical support for the multi-window multi-pathway model for humans includes different effects of different kinds of brain damage.

6.2.9 Motor pathways

Connections from central to motor mechanisms may use either the ‘peephole’ model, with all motor signals going through a narrow channel from the central system (e.g. bottom right as in the Omega model), or a ‘multi-window’ architecture where different sorts of instructions from different central layers can go to a layered, hierarchical motor system, which performs the necessary decomposition to low level motor signals along with integration as needed, as in (Albus, 1981) and fig. 2(b). The latter seems to be required for skilled performance of complex hierarchical actions.

6.2.10 Specialised ‘boxes’ *vs* emergence

Some architecture diagrams have a box labelled ‘emotions’. In others, emotions, like ‘thrashing’ in an operating system. are treated as emergent properties of interactions between functional components such as alarm mechanisms, motive generators and attention filters, (Wright et al., 1996; Sloman, 2001a). An architecture like fig. 2(b) can explain at least three different classes of emergent emotions involving disturbances caused by or affecting different layers of the architecture. Whether a capability needs a component, or emergent interactions between components is not always clear. The attention filter in fig. 2(b) could use either a special mechanism (easier to implement and control) or the emergent effects of interactions between competing components (more general and flexible) although the trade-offs depend on the particular architecture. The ‘emergent’ approach is illustrated by the contention scheduling model.

6.2.11 Dependence on external language

Some models postulate a close link between high level internal processes and an external language. For instance, some claim that mechanisms like meta-management require a public language and social system, and some regard language as essential for human-like minds (Dennett, 1996). Others (Sloman, 1979) regard internal mechanisms and formalisms for deliberation and high level self-evaluation as pre-cursors to the development of human language as we know it. (Compare Barkley (Barkley, 1997)). It appears from the capabilities of many animals, that rich and complex information processing mechanisms evolved long before external human-like languages, and probably still underpin them. In that sense the use of ‘language’ to think with is prior to its use in external communication, though we are not denying the impact of external language.

6.2.12 Internal *vs* partly external implementation

Most AI design work focuses on internal processing. However, Simon pointed out in 1969 that animals often use the environment as a short term or long term memory: so their implementation extends beyond their bodies. Human examples include trail-blazing and calculating on paper. Strawson argued in (Strawson, 1959) that what is *within* an individual cannot *suffice* to determine that some internal representation or thought refers to the Eiffel tower, as opposed to an exactly similar object on a ‘twin earth’. Unique reference depends in part on the causal and spatial relationships to the thing referred to. So not *all* aspects of human-like thought can be fully implemented internally: some depend on external relations (Sloman, 1985; Sloman, 1987).

6.2.13 Self-bootstrapped ontologies

We have argued that if we specify an architecture we shall understand what sorts of processes can occur in it, and will be able to define an appropriate set of concepts for describing its 'mental' states.

However, some learning mechanisms can develop their own ways of clustering phenomena according to what they have been exposed to, and their successes and failures. In a robot with the architecture in fig. 2(b) the meta-management layer might develop a collection of concepts for categorising its own internal states and processes that nobody else can understand intuitively because nobody else has been through that particular history of learning processes. Subsequent effects of using those 'personal' concepts may exacerbate the complexity and idiosyncratic character of the robot's internal processing. (Compare the difficulty of understanding what a complex neural network is doing, after it has been trained.) Partial understanding 'from outside' might come from analysing the history and its effects on the architecture. For systems with that degree of sophistication and reflective capability, scientific understanding of their processing may forever be limited to very coarse-grained categorisations and generalisations.

7 Discussion

In this short paper we have tried to show (albeit with much missing detail) that a general formulation of a wide class of architectures can facilitate comparative analysis of different proposed architectures, by providing a common vocabulary for describing structure and function. Our CogAff schema is a first-draft example that accommodates a wide range of architectures (though not, for instance, distributed software systems). We have tried to bring out some of the options that may need to be considered when trying to design, compare and evaluate architectures, though we have said nothing here about the even larger variety of multi-agent architectures.

After specifying a particular case of the schema, we can analyse the types of capabilities, states and processes enabled within that special case. This provides a basis for refining vague or indeterminate cluster concepts of ordinary language (e.g. 'emotion', 'believe', 'intention', 'learning') so that they become more precise, with clear criteria for deciding which animals or robots exemplify them. This avoids endless debates about which animals 'really' think, etc.

Different architectures will support different collections of concepts, and care is required if familiar human mental concepts are being used: they may not always be applicable to some of the simpler artificial systems, illustrating McDermott's argument.

A schema such as CogAff also provides a basis for developing an enriched theory of learning where varieties of learning and development that are possible depend not only on the mechanisms that are present within components of the architecture, but also on the scope for the architecture to extend itself with new components or new links. Because so many types of change are possible in more complex systems, we can expect to have to replace our ordinary concepts of 'learning' and 'development' with a family of more precise architecture-based concepts. (There is no space here for a full analysis.)

We can use the schema to explore some of the varieties of evolutionary trajectories. In some recent experiments (Scheutz and Logan, 2001) it appears that for simple sorts of reactive agents and a range of environments, adding simple affective mechanisms is more beneficial (for survival over many generations) than adding (simplified) deliberative capabilities. Because a schema like CogAff invites us to consider ways of extending an architecture which does not already have all possible links and components, we can use it to define 'neighbourhoods' in design space. We can then explore those neighbourhoods analytically or by doing computational experiments, or by looking for paleontological evidence.

Further investigation might help us understand better why the vast majority of the earth's biomass consists of relatively unintelligent organisms, with only reactive components. Part of the answer may be requirements of food chains needed to support animals with more complex brains! However, there may be more fundamental reasons why large numbers of relatively stupid, but inexpensive and expendable, individuals (with affective control states) are normally more successful than smaller numbers of larger, more expensive and more intelligent organisms. By understanding those reasons we can understand the exceptional conditions that promote evolution of additional, more expensive, deliberative mechanisms.

In later stages of evolution, the architecture might support new types of interaction and the development of a culture. For instance if the meta-management layer, which monitors, categorises, evaluates and to some extent controls or redirects other parts of the system, absorbs many of its categories and its strategies from the culture, then the same concepts can be used both for self-description and for other-description: a form of social control.

Versions of the third layer providing the ability to attend to and reflect on some intermediate perceptual processes could cause intelligent robots to discover *qualia*, and wonder whether humans have them!

We can also use our framework to clarify and refine architectural concepts developed in psychology. The common reference to 'executive function' by psychologists and brain scientists conflates aspects of the deliberative and meta-management layers. That they are different is shown by the existence of AI systems with sophisticated planning and problem solving and plan-execution capabilities, but without meta-management (reflective) capabilities. In consequence, some planners cannot notice obvious types of redundancy in the plans they produce, nor subtle looping behaviour when planning. After developing these ideas we found that the neuropsychiatrist Barkley (*op. cit.*) had reached closely related conclusions starting from empirical data.

Study of a general schema for a wide class of architectures should help AI researchers designing and comparing agent architectures, and also philosophers, brain scientists, social scientists, ethologists and evolutionary biologists. CogAff seems to be a useful first draft, though much remains to be done.

Notes and Acknowledgements

This work benefited from discussions with Margaret Boden, Brian Logan, Marvin Minsky, Matthias Scheutz, Ron Chrisley, Push Singh, and many colleagues and students in the Cognition and Affect project at The University of Birmingham and elsewhere.

I owe much to early encouragement from Margaret Boden, who was responsible for publication in 1978 of *The Computer Revolution in Philosophy*. But for her it would never have been completed.

Our recent work was supported by the Leverhulme Trust.

Previous support has come from GEC Research Laboratories, The Renaissance Trust (Gerry Martin), the UK Joint Research Council Initiative in HCI and Cognitive Science, The University of Birmingham, and DERA.

Our papers can be found at
http://www.cs.bham.ac.uk/research/cogaff/

and our tools at
http://www.cs.bham.ac.uk/research/poplog/freepoplog.html
http://www.cs.bham.ac.uk/~axs/cogaff/simagent.html

There are several (overlapping) presentations on topics discussed here:
http://www.cs.bham.ac.uk/research/cogaff/talks/

References

Albus, J. (1981). *Brains, Behaviour and Robotics*. Byte Books, McGraw Hill, Peterborough, N.H.

Alechina, N. and Logan, B. (2001). State space search with prioritised soft constraints. *Applied Intelligence*, 14(3):263–272.

Arbib, M. A. (2002). From Rana Computatrix to Homo Loquens: A computational neuroethology of language evolution. In Damper, R. I. et al., editors, *WGW'02 Biologically-inspired robotics: The legacy of W. Grey Walter*, pages 12–31, Bristol. Hewlett Packard Research Labs.

Barkley, R. A. (1997). *ADHD and the nature of self-control*. The Guildford Press, New York.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Cooper, R. and Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4):297–338.

Craik, K. (1943). *The Nature of Explanation*. Cambridge University Press, London, New York.

Damasio, A. (1994). *Descartes' Error, Emotion Reason and the Human Brain*. Grosset/Putnam Books, New York.

Dennett, D. (1996). *Kinds of minds: towards an understanding of consciousness*. Weidenfeld and Nicholson, London.

Dennett, D. C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, MA.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA.

Johnson-Laird, P. (1988). *The Computer and the Mind: An Introduction to Cognitive Science*. Fontana Press, London. (Second edn. 1993).

Kant, I. (1781). *Critique of Pure Reason*. Macmillan, London. Translated (1929) by Norman Kemp Smith.

Kennedy, C. M. and Sloman, A. (2003). Autonomous recovery from hostile code insertion using distributed reflection. *Journal of Cognitive Systems Research*, 4(2):89–117.

Lees, M., Logan, B., Oguara, T., and Theodoropoulos, G. (2003a). Simulating Agent-Based Systems with HLA: The case of SIM_AGENT – Part II. In *Proceedings of the 2003 European Simulation Interoperability Workshop*.

Lees, M., Logan, B., and Theodoropoulos, G. (2003b). Adaptive Optimistic Synchronisation for Multi-Agent Simulation. In *Proceedings of the 17th European Simulation Multiconference (ESM 2003)*.

Lees, M., Logan, B., and Theodoropoulos, G. ((to appear)). Agents, Computer Games and HLA. *International Journal of Simulation Systems, Science and Technology*.

Logan, B. and Alechina, N. (1998). A* with bounded costs. In *Proceedings of the 15th National Conference on Artificial Intelligence–AAAI-98*. Also Birmingham School of Computer Science technical report CSRP-98-09.

McDermott, D. (1981). Artificial intelligence meets natural stupidity. In Haugeland, J., editor, *Mind Design*. MIT Press, Cambridge, MA.

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

Nagel, T. (1981). What is it like to be a bat. In Hofstadter, D. and D.C.Dennett, editors, *The mind's I: Fantasies and Reflections on Self and Soul*, pages 391–403. Penguin Books.

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

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

Nilsson, N. (1994). Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research*, 1:139–158.

Nilsson, N. (1998). *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, San Francisco.

Picard, R. (1997). *Affective Computing*. MIT Press, Cambridge, MA; London, England.

Popper, K. (1976). *Unended Quest*. Fontana/Collins, Glasgow.

Ritter, F., editor (2002). *Techniques for Modeling Human Performance in Synthetic Environments: A Supplementary Review*. Defense Technical Information Center, Ft. Belvoir, VA. http://ritter.ist.psu.edu/papers/SOAR-Jun03.pdf.

Scheutz, M. (2002). Agents with or without emotions? In *Proceedings FLAIRS 02*, pages 89–94. AAAI Press.

Scheutz, M. and Logan, B. (2001). Affective vs. deliberative agent control. In C. Johnson, *et al.*., editor, *Proceedings Symposium on Emotion, cognition and affective computing AISB01 Convention*, York.

Scheutz, M. and Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In Ning Zhong, *et al.*, editor, *Intelligent Agent Technology: Research and Development*, pages 200–209. World Scientific Publisher, New Jersey.

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

Sloman, A. (1969). How to derive "better" from "is". *American Phil. Quarterly*, 6:43–52. http://www.cs.bham.ac.uk/research/cogaff/sloman.better.html.

Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, pages 209–226, London. William Kaufmann. http://www.cs.bham.ac.uk/research/cogaff/04.html#200407.

Sloman, A. (1978). *The Computer Revolution in Philosophy.* Harvester Press (and Humanities Press), Hassocks, Sussex. http://www.cs.bham.ac.uk/research/cogaff/crp.

Sloman, A. (1979). The primacy of non-communicative language. In MacCafferty, M. and Gray, K., editors, *The analysis of Meaning: Informatics 5 Proceedings ASLIB/BCS Conference, Oxford, March 1979*, pages 1–15, London. Aslib. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#43.

Sloman, A. (1985). What enables a machine to understand? In *Proc 9th IJCAI*, pages 995–1001, Los Angeles.

Sloman, A. (1987). Reference without causal links. In du Boulay, J., D.Hogg, and L.Steels, editors, *Advances in Artificial Intelligence - II*, pages 369–381. North Holland, Dordrecht.

Sloman, A. (1989). On designing a visual system (towards a gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337. http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7.

Sloman, A. (1994). Explorations in design space. In Cohn, A., editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester. John Wiley.

Sloman, A. (1996). Actual possibilities. In Aiello, L. and Shapiro, S., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, pages 627–638, Boston, MA. Morgan Kaufmann Publishers.

Sloman, A. (2000a). Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Dautenhahn, K., editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam.

Sloman, A. (2000b). Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In M.Schoenauer, *et al.*., editor, *Parallel Problem Solving from Nature – PPSN VI*, Lecture Notes in Computer Science, No 1917, pages 3–16, Berlin. Springer-Verlag.

Sloman, A. (2000c). Models of models of mind. In Lee, M., editor, *Proceedings of Symposium on How to Design a Functioning Mind, AISB'00*, pages 1–9, Birmingham. AISB.

Sloman, A. (2001a). Beyond shallow models of emotion. *Cognitive Processing: International Quarterly of Cognitive Science*, 2(1):177–198.

Sloman, A. (2001b). Evolvable biologically plausible visual architectures. In Cootes, T. and Taylor, C., editors, *Proceedings of British Machine Vision Conference*, pages 313–322, Manchester. BMVA.

Sloman, A. (2001c). Varieties of Affect and the CogAff Architecture Schema. In Johnson, C., editor, *Proceedings Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention*, pages 39–48, York.

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, 10(4-5):113–172.

Sloman, A., Chrisley, R., and Scheutz, M. (2005). The architectural basis of affective states and processes. In Arbib, M. and Fellous, J.-M., editors, *Who Needs Emotions?: The Brain Meets the Robot*, pages 203–244. Oxford University Press, New York. http://www.cs.bham.ac.uk/research/cogaff/03.html#200305.

Sloman, A. and Chrisley, R. L. (2005). More things than are dreamt of in your biology: Information-processing in biologically-inspired robots. *Cognitive Systems Research*, 6(2):145–174.

Sloman, A. and Logan, B. (1999). Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association for Computing Machinery*, 42(3):71–77. http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#49.

Sloman, A. and Logan, B. (2000). Evolvable architectures for human-like minds. In Hatano, G., Okada, N., and Tanabe, H., editors, *Affective Minds*, pages 169–181. Elsevier, Amsterdam. Invited talk at Toyota Conference 1999.

Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. Methuen, London.

Sussman, G. (1975). *A computational model of skill acquisition*. American Elsevier.

Weiskrantz, L. (1997). *Consciousness Lost and Found*. Oxford University Press, New York, Oxford.

Wright, I., Sloman, A., and Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126. http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#2.