A Framework for Comparing Agent Architectures

Aaron Sloman
School of Computer Science
University of Birmingham
Birmingham
B15 2TT. England
a.sloman@cs.bham.ac.uk

Matthias Scheutz
Department of Computer Science and Engineering
351 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA
mscheutz@cse.nd.edu

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 representation of a general 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 providing a framework for evaluating alternative architectures. As a first-draft example we present the CogAff architecture schema, and show how it provides a draft framework. But there is much still to be done.

1 Introduction

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.

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 (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 (1943) proposed that animals build 'models' of reality in order to explore possible actions safely without actually performing them. Popper (in his 1976 and earlier works) advocated similar mechanisms allowing our mistaken hypotheses to 'die' instead of us. Recent work has added more detail. Albus (1981, p.184) 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 [2], though most psychologists and neuroscientists find it very difficult to think about virtual machine architectures. Shallice and Cooper are among the exceptions [4].

In the meantime, AI researchers have been exploring many sorts of architectures. See Nilsson's (1998, Ch 25) account of *triple tower* and *triple layer* models. Architectures like SOAR, ACT-R, and Minsky's *Society of Mind* have inspired many researchers, but there is 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 makes a good start. In short, there are no adequate surveys of 'design space' and 'niche space' and their relationships [24]. As a partial remedy, we offer the CogAff schema depicted in figures 1(a), (b) and 2(a), (b), and described below.

2 Conceptual confusions

A problem surrounding the study of architectures is the diversity of high level aims of AI researchers. Some are merely trying to solve engineering problems and care only whether their solutions work well. Others are explicitly attempting 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

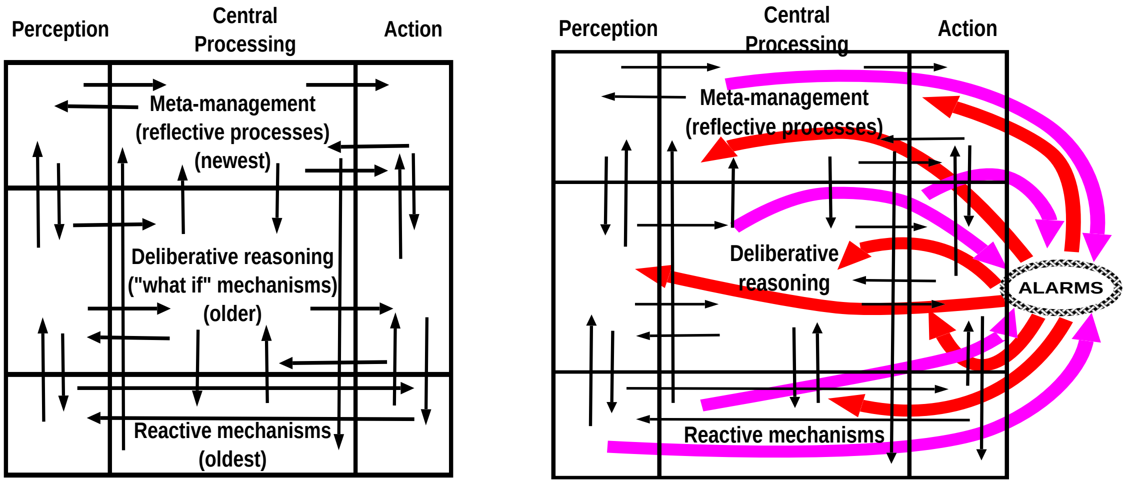

Figure 1: (a) (b)

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. 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. Humans seem to have all three. 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.

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' [8, 22, 26].

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 [28, 27, 16] Social scientists tend to define 'emotion' so as to focus on social phenomena, such as embarrassment, attachments, guilt or pride, whereas a brain scientist might define it to refer to brain processes and widespread animal behaviours. It has dozens of definitions in the psychological and philosophical literature, because different authors attend to different subsets of emotional phenomena.

McDermott's (1981) critique 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 [25, 22, 26].

2.1 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. [6, 14, 23, 28]. Likewise, different concepts of 'seeing' relate to visual pathways through different subsystems in a larger architecture. 'Blindsight' [31] is explained by damage to connections between meta-management and intermediate high level perceptual buffers, 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' [12] may be no more obscure to us than 'What it is like to be a baby', or an Altzheimer's sufferer.

2.2 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, …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 them. Worse, we may be unaware of the full range 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 as we have defined primary, secondary and tertiary emotions in terms of reactive, deliberative and meta-management mechanisms e.g. [25].

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.

3 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 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.

Different subsets of the CogAff schema can be used in different architectures. A conjectured human-like schema H-CogAff,[1] is depicted crudely in fig. 2. Much simpler instances are also accommodated. E.g. 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 [30] combined portions of the top two layers. SOAR's 'impasse detection' is a type of meta-management. Brooks (1991) subsumption architectures include multiple control levels all within the reactive layer, and nothing in the other layers.

3.1 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 [3], 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. This, and other ambiguities, may be reduced if people proposing architectures can at least agree on a broad framework specifying a class of architectures and terminology for describing and comparing them, as illustrated in the next section.

3.2 Dimensions of architectural variation

We present a first draft list of dimensions in which architectures can be compared.

1. Pipelined vs *concurrently active layers*

Often [13] the layers have a sequential processing function: sensory information comes in via low level sensors ('bottom left'), gets abstracted as it goes up

1. 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.

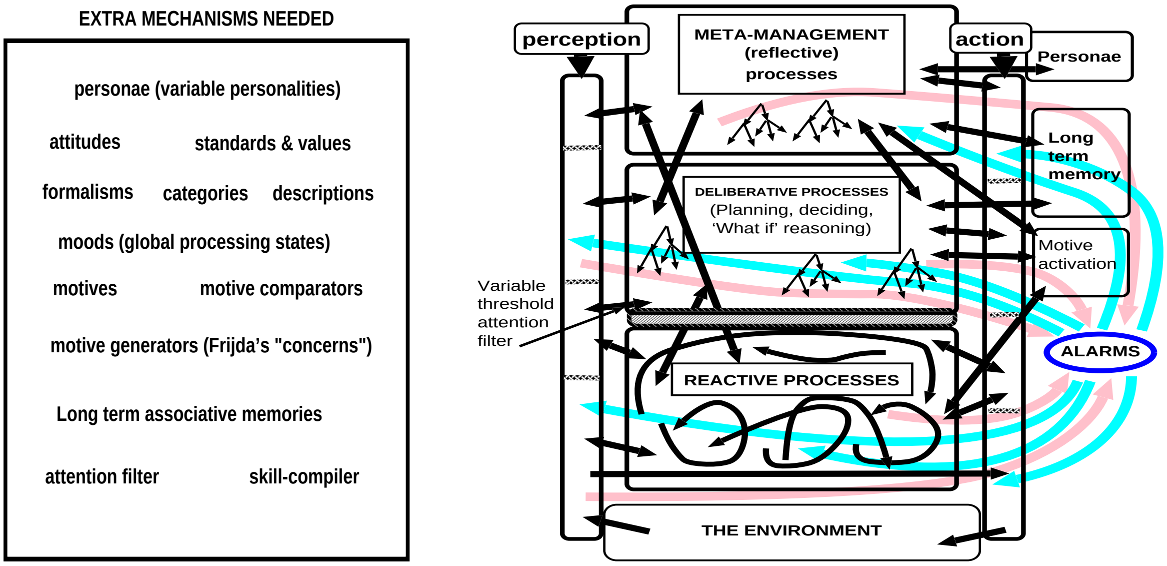

Figure 2: (a) (b)
The 'Human-like' sub-schema H-Cogaff. (a) lists some components supporting motive processing and 'what if' reasoning in deliberative and meta-management layers. Humans seem to have all those mechanisms, perhaps organised as in (b), where the alarm system might correspond to the brain's 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.

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. [1]. The 'contention scheduling' model [4] 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).

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.

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 a 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.

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, and operate at 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 possible future actions and consequences of actions, to categorise them, evaluate them, and make selections. This

could support planning, explanation of past events, mathematical reasoning, and general counterfactual reasoning. Such deliberative mechanisms require temporary workspace containing changing structures. These are not needed for most reactive systems.

Of course, the deliberative mechanisms which go beyond reactive mechanisms in explicitly representing alternative actions prior to selection might themselves 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. to reduce time complexity. then these mechanisms could also be accommodated.

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. 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 to either reactive behaviours as in athletic training, or to abstract concepts, as in learning to read a foreign language or detect a style of painting.[2]

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.

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 centrally in the reactive layer, some might be detected and resolved using symbolic reasoning in the reactive 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 architectures use 'numerical' conflict resolution, e.g. voting mechanisms, others rule-based or problem-solving decision systems capable of creative compromises, and some hybrid mixtures.

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 [22, 28] 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 subsystems (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. Differential effects of different kinds

2. 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.

of brain damage seem to support the multi-window multi-pathway model for humans.

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 [1] and fig. 2(b). The latter seems to be required for skilled performance of complex hierarchical actions.

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, [32, 25]. 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 effects of interactions between competing components (more general and flexible) although the trad-eoffs depend on the particular architecture. The latter approach is found in the contention scheduling model.

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 [7]. Others [19] 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 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.

12. Internal vs *partly external implementation*
Most AI design work focuses on internal processing. However, Simon (1969) pointed out that animals often use the environment as a short term memory: so their implementation extends beyond their bodies. Human examples include trail-blazing and calculating on paper. Strawson (1959) pointed out 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 [20, 21].

13. Self-bootstrapped ontologies
We have been arguing that when we have specified 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.

4 Discussion

We have tried to show (albeit with much missing detail, on account of space limitations) 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 possibly not all, especially totally unstructured architectures). 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 and clear criteria for deciding which animals or robots exemplify them can replace endless debates.

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 [17] 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 general reasons why sometimes large numbers of relatively stupid, but inexpensive and expendable individuals (with affective control states) are more effective than smaller numbers of larger, more expensive and more intelligent organisms.

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 for self-description and for other-description: a form of social control.

The third layer also provides the ability to attend to and reflect on one's own mental states, which could cause intelligent robots to discover qualia, and wonder whether humans have them.

We can use our framework to clarify some scientific concepts. The common reference to 'executive function' by psychologists and brain scientists seems to conflate aspects of the deliberative layer and aspects of the meta-management layer. 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. A symptom would be a planner that doesn't notice an obvious type of redundancy in the plan it produces, or subtle looping behaviour. After developing these ideas we found that the neuropsychiatrist Barkley (1997) 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 has benefited from discussions with Brian Logan, Marvin Minsky, and many colleagues and students in the Cognition and Affect project at The University of Birmingham and elsewhere. The work is supported by the Leverhulme Trust.
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

References

[1] J. Albus. *Brains, Behaviour and Robotics*. Byte Books, McGraw Hill, Peterborough, N.H., 1981.

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

[3] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

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

[5] K. Craik. *The Nature of Explanation*. Cambridge University Press, London, New York, 1943.

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

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

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

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

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

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

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

[13] N. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, San Francisco, 1998.

[14] R. Picard. *Affective Computing*. MIT Press, Cambridge, Mass, London, England, 1997.

[15] K. Popper. *Unended Quest*. Fontana/Collins, Glasgow, 1976.

[16] M. Scheutz. Agents With or Without Emotions? In *Proceedings FLAIRS 02*. AAAI Press, 2002.

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

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

[19] A. Sloman. The primacy of non-communicative language. In M. MacCafferty and K. Gray, editors, *The analysis of Meaning: Informatics 5 Proceedings ASLIB/BCS Conference, Oxford, March 1979*, pages 1–15, London, 1979. Aslib.

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

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

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

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

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

[25] A. Sloman. Beyond Shallow Models of Emotion. *Cognitive Processing: International Quarterly of Cognitive Science*, 2(1):177–198, 2001.

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

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

[28] A. Sloman and B. Logan. Evolvable architectures for human-like minds. In G. Hatano, N. Okada, and H. Tanabe, editors, *Affective Minds*, pages 169–181. Elsevier, Amsterdam, 2000.

[29] P. F. Strawson. *Individuals: An essay in descriptive metaphysics*. Methuen, London, 1959.

[30] G. Sussman. *A Computational Model of Skill Acquisition*. American Elsevier, 1975.

[31] L. Weiskrantz. *Consciousness Lost and Found.* Oxford University Press, New York, Oxford, 1997.

[32] I. Wright, A. Sloman, and L. Beaudoin. Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101–126, 1996. Repr. in R.L.Chrisley (Ed.), *Artificial Intelligence: Critical Concepts in Cognitive Science*, Vol IV, Routledge, London, 2000.